

# MINA: A Multilingual LLM-Powered Legal Assistant Agent for Empowering Access to Justice in Bangladesh

Anonymous ACL submission

## Abstract

Bangladesh’s low-income population faces major barriers to affordable legal advice due to complex legal language, procedural opacity, and high costs. Existing AI legal assistants lack Bengali-language support and jurisdiction-specific adaptation, limiting their effectiveness. To address this, we developed MINA, a multilingual LLM-based legal assistant tailored for the Bangladeshi context. It employs multilingual embeddings and a RAG-based chain-of-tools framework for retrieval, reasoning, translation, and document generation, delivering context-aware legal drafts, citations, and plain-language explanations via an interactive chat interface. Evaluated by law faculty from leading Bangladeshi universities across all stages of the 2022 and 2023 Bangladesh Bar Council examinations, MINA achieved scores of 75–80% in the preliminary MCQs, written, and simulated viva voce components. These results matched or surpassed average human performance, demonstrating strong clarity, contextual understanding, and sound legal reasoning, while operating at approximately 0.1-0.6% of the cost of human lawyers. These results confirm its potential as a low-cost, multilingual AI assistant that automates key legal tasks and scales access to justice, offering a real-world details on building domain-specific, low-resource systems and addressing challenges of multilingual adaptation, efficiency, and sustainable public-service AI deployment.

## 1 Introduction

Access to affordable legal advice remains a major challenge for economically disadvantaged groups in Bangladesh, largely due to complex legal language, opaque procedures, and high costs (Knock et al., 2024; Raj, 2024; Islam et al., 2024; Akter, 2017; Wasi et al., 2024). This contributes to a severe access to justice crisis that disproportionately affects marginalized populations. The judiciary is burdened by an enormous backlog of 3.7 to 4.4

million cases and operates with only 2,100 judges (Yesmin, 2025; Ahmed, 2021), translating to one judge per 90,000 people and resulting in delays of 10 to 60 years (Jaan, 2023). Civil disputes often drag on for decades, while criminal cases like the Chittagong Arms Haul have remained unresolved for over 15 years (Knock et al., 2024; Khalilullah, 2025). Outdated procedural laws enable frequent adjournments and exploitation by lawyers, further slowing case resolution. Meanwhile, unregulated lawyer fees make legal representation unaffordable for most (Jaan, 2023), and public legal aid faces constraints from limited funding, strict eligibility, and engagement (Tahura and Alam, 2025).

The absence of digital infrastructure and automation intensifies inefficiencies and delays, compounded by manual court processes. Additionally, widespread legal illiteracy, fear of retaliation, and dependence on biased informal mechanisms discourage many from seeking justice, perpetuating inequality and eroding trust in the system (Knock et al., 2024; Islam et al., 2024; Wasi et al., 2024). Consequently, many avoid formal legal engagement altogether, deepening existing disparities and weakening the rule of law (Knock et al., 2024). Natural Language Processing (NLP) tools can help bridge this gap by simplifying complex legal texts, automating document analysis, and providing accessible guidance to under-served populations.

Motivated by the urgent need to improve legal access for Bangladesh’s low-income population, who face significant barriers due to complex legal language, long procedures, and high costs, we developed MINA, a multilingual LLM-based legal assistant tailored to the Bangladeshi legal context. Unlike prior systems focused on English and Western legal frameworks, our assistant integrates Bengali and English support using multilingual embeddings within a Retrieval-Augmented Generation (RAG) framework that combines retrieval, reasoning, drafting, translation, and citation inser-

043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083

tion. Delivered via an interactive chat interface, it assists users in drafting petitions, referencing statutes, and simplifying complex legal language into plain Bengali.

We evaluated our agentic system MINA rigorously by law faculty experts across all stages of the 2022 and 2023 Bangladesh Bar Council Examinations, including the Preliminary multiple-choice questions, Written exam, and Viva Voce oral exam (conducted via chat UI). Consistently scoring 75–80%, it matched or exceeded average human performance, passing both years of the Bar Council exam and demonstrating robust legal reasoning and contextual understanding.

Overall, our contributions include: (i) *the development of a localized multilingual legal assistant with integrated RAG, reasoning, drafting, translation, and citation capabilities*; (ii) *rigorous evaluation demonstrating human-comparable performance, robustness, and generalization across exam years, modalities, and diverse legal tasks*; and (iii) *passing the Bangladesh Bar Council exam, complemented by a comprehensive error analysis and insights for deploying scalable, cost-effective legal assistance to under-served populations*.

While our system uses established components like multilingual embeddings, RAG, and LangGraph-style agents, its novelty lies in adapting them to a bilingual, low-resource legal environment. The two-stage RAG pipeline retrieves statutes at the Act and Section levels, preventing conflation of unrelated provisions and ensuring coherent legal responses. Multilingual embeddings and a legal dictionary support interpretation of colonial-era and Farsi-influenced terminology, with external tools used selectively to augment context. This system-level design demonstrates practical deployment for accessible legal assistance, validated rigorously across multiple evaluation stages.

## 2 Preliminaries

Legal NLP is transforming legal practice by automating document analysis, contract review, and research, significantly reducing time and costs (Lai et al., 2023; Yan, 2023; Frankenreiter and Nyarko, 2023; Zhong et al., 2020), with transformer models like BERT and GPT enhancing contextual understanding for more accurate insights (Martin et al., 2024; Jiang et al., 2024). Large Language Models (LLMs) extend this potential by simplifying complex legal language, translating statutes and proce-

dures into plain Bengali, and automating repetitive tasks such as document drafting and legal research, which is critical in a system where formal aid is underfunded and unaffordable for many (Safdie, 2025; Jaan, 2023). Multilingual LLM-powered agents capable of processing both Bengali and English legal texts offer accurate cross-lingual retrieval and can handle mixed-language documents prevalent in Bangladesh, thereby enhancing accessibility and efficiency.

However, existing Legal NLP tools remain inadequate for Bangladesh due to linguistic, legal, and socio-economic complexities: Bengali support suffers from scarce annotated datasets, limited digital content, and underdeveloped tools for tasks like tokenization and NER (Wasi et al., 2024), while English-centric models struggle with Bengali’s complex morphology and script. Bangladesh follows the Common Law system (a legacy of British colonial rule). The judiciary operates on an adversarial basis, relying heavily on codified statutes (like the Code of Civil Procedure, 1908 and Penal Code, 1860) and judicial precedent. Our evaluation via the Bar Council Exam is specifically designed to test proficiency within this Common Law framework. Furthermore, Bangladesh’s legal system, rooted in colonial-era codes and Farsi-influenced terminology, poses additional challenges, as models lack jurisdiction-specific knowledge and cannot effectively interpret archaic legal Bengali (Mizan, 2021; Asjad, 2023). Compounding these technical issues are gendered digital divides, poor translation quality (Asjad, 2023), low digital literacy, and limited offline legal resources, underscoring the urgent need for a localized, context-aware AI solution tailored to Bangladesh’s unique legal environment. Such an AI system could democratize legal assistance, empower marginalized populations, and help alleviate the country’s severe access-to-justice crisis.

## 3 Methodology: MINA

Our system is anchored by an Orchestrator Agent that evaluates user input, chat history, and documents to select the appropriate response pathway. When internal context is insufficient, a two-stage RAG pipeline retrieves relevant Acts and Sections using Cohere-generated keywords and multilingual embeddings over Chroma vector stores of Bangladeshi laws. Retrieved texts inform final reasoning, while external tools (web search, parsers)

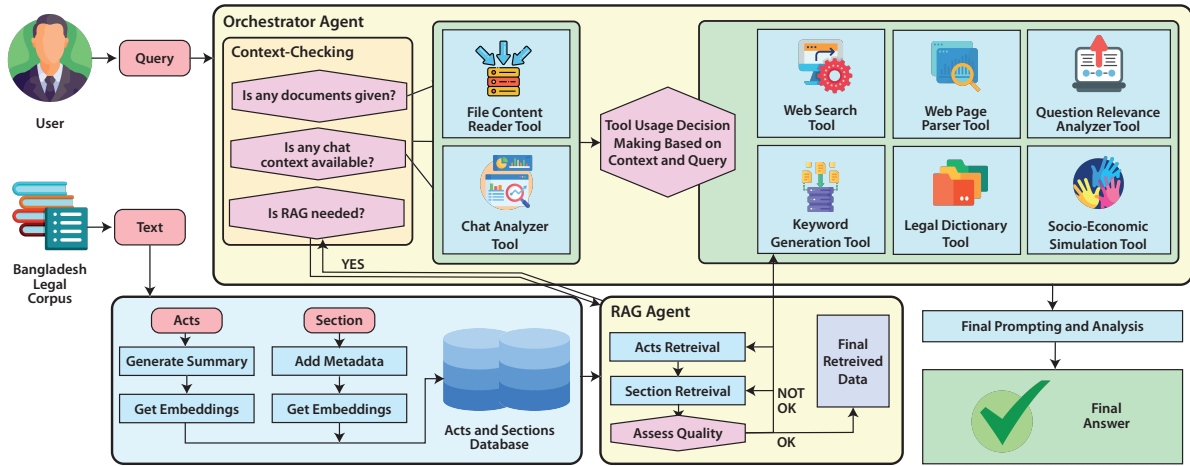


Figure 1: System Architecture and Workflow of our Multilingual Legal Assistant Agent for Bangladesh

are used only when necessary. The system handles both Bengali and English, interprets colonial-era and Farsi-influenced terms via a custom legal dictionary, and includes a socio-economic simulation module. Optimized for low-resource settings, it provides context-aware, jurisdiction-specific legal assistance via a simple chat interface, simplifying legal language and automating core tasks to empower users without legal training or access to affordable services.

### 3.1 Architecture Design

This section details each component of the architecture, their interactions, and how they collectively address the core legal accessibility challenges in the region, as outlined in Figure 1.

#### 3.1.1 Retrieval-Augmented Generation (RAG)

Our two-stage RAG subsystem retrieves legal texts and enhances response relevance through sequential Act and Section retrieval, using dedicated vectorDBs built from Bangladeshi legal corpora.

**Data.** In the Bangladeshi legal system, Acts are the primary legislative units, each containing an average of 24 Sections. Act titles average 50.30 characters, Section titles 38.07 characters, and Section content 736.69 characters, reflecting their greater interpretive depth. As per April 2025, the dataset contains 595 Acts comprising 18, 023 Sections (avg. 24 sections/act), with average act name and detail lengths of 50 and 438 characters, respectively, and average section name and detail lengths of 38 and 736 characters.

**Database Development.** Initial experiments showed that naïve retrieval often produced responses by combining content from unrelated Acts. To address this, we developed two separate vector

databases. The Act database was constructed by collecting full-text legislative data from the official Bangladesh Law and Justice website. For each Act, we generated an LLM-based summary to create a compact, searchable representation, indexed alongside metadata. The Section database was built by chunking individual sections (1–2 chunks each) and attaching relevant Act metadata for precise filtering during Section-level search.

**RAG Pipeline and Workflow.** When a query is received, the Keyword Generator tool produces semantic keywords to retrieve the top  $n_{acts}$  (default = 5) most relevant Acts from indexed summaries. These Act IDs then filter the Section database to find the most relevant  $n_{sections}$  (default = 10) legal provisions. Multilingual embeddings ensure accurate retrieval across Bengali and English texts. Retrieved sections are checked for contextual relevance; if insufficient, the query is refined and rerun. Once validated, the Acts and Sections are sent to the Orchestrator Agent for final reasoning. This two-step process ensures broad coverage at the Act level and precise grounding at the Section level.

#### 3.1.2 Agents

Our system is powered by two core agents: the *Orchestrator Agent* and the *RAG Agent*, each with distinct roles in the workflow.

**Orchestrator Agent** serves as the central decision-maker, evaluating user queries in context, considering prior conversation turns and any uploaded documents, to determine if a direct response can be generated. If the context is sufficient, it proceeds to answer generation. Otherwise, it delegates the task to the RAG Agent.

**RAG Agent** manages the retrieval process end-to-end, from keyword generation to assessing the

255	relevance and jurisdictional appropriateness of re-	<b>3.1.4 LLM Integration and Prompts</b>	303
256	trieved legal materials. This separation of roles	At the core of our system’s understanding	304
257	enhances system modularity and simplifies mainte-	layer is a multilingual LLM. In our experi-	305
258	nance, debugging, and extensibility.	ments, we evaluate a broad spectrum of mod-	306
259	Both agents operate within a <code>LangGraph</code> -based	els spanning both proprietary and open-source	307
260	state machine, enabling persistent memory across	families, including OpenAI’s <i>GPT-4o</i> (OpenAI	308
261	interactions. This design supports complex, multi-	et al., 2024), Google’s <i>gemini-2.0-flash</i> (propri-	309
262	turn legal consultations where reasoning must	etary), <i>gemini-2.5-flash</i> (proprietary) (Team et al.,	310
263	evolve. The agentic structure also enables condi-	2025a), <i>gemma-3-4b-it</i> , <i>gemma-3-12b-it</i> , <i>gemma-</i>	311
264	tional execution, allowing the system to adapt	<i>3-27b-it</i> (Team et al., 2025b), Meta’s <i>llama3.2-</i>	312
265	its workflow dynamically without compromising	<i>1b-instruct</i> , <i>llama3.2-3b-instruct</i> , <i>llama3.1-8b-</i>	313
266	performance or legal integrity.	<i>instruct</i> , <i>llama3.1-70b-instruct</i> (Grattafiori et al.,	314
267		2024), Cohere’s <i>command-a-03-2025</i> (Cohere	315
268		et al., 2025), Qwen’s <i>qwen3-4b-instruct-2507</i> , and	316
269		<i>qwen3-30b-a3b-instruct-2507</i> (Yang et al., 2025).	317
270		All models were evaluated across system modules	318
271		for legal drafting, question answering, and prompt	319
272		compliance.	320
273		For retrieval, we employ Cohere’s <i>embed-</i>	321
274		<i>multilingual-light-v3.0</i> , which supports semantic	322
275		similarity in both Bengali and English, crucial for	323
276		processing bilingual legal content in Bangladesh.	324
277		Prompt engineering ensures legal accuracy and ju-	325
278		risdictional relevance. Prompts are constructed us-	326
279		ing prior conversation turns, uploaded documents,	327
280		and the RAG state. All prompts are detailed in	328
281		Appendix L.	329
282			
283		<b>3.2 Adaptive Agent Workflow</b>	330
284		Our system operates through two primary working	331
285		modes: <i>situational</i> and <i>decisional</i> . In the <b>situa-</b>	332
286		<b>tional</b> pathway, if a user uploads a document, it is	333
287		routed through the Document Analyzer, which ex-	334
288		tracts and prepares relevant legal content for down-	335
289		stream use. Alternatively, when prior user interac-	336
290		tions are available, the Chat Analyzer identifies re-	337
291		lated past queries to provide contextual grounding	338
292		for the current request. In the <b>decisional</b> pathway,	339
293		the system assesses whether retrieval-augmented	340
294		generation (RAG) is necessary. When internal con-	341
295		text, such as documents or chat history, is insuf-	342
296		ficient or absent, we trigger the RAG workflow.	343
297		This involves generating semantic keywords, re-	344
298		trieving relevant vectors from pre-indexed legal	345
299		corpora (e.g., Bangladeshi statutes and sections),	346
300		and composing a contextually grounded response.	347
301		This flexible routing mechanism enables the agent	348
302		to adapt intelligently based on the type of input,	349
		historical interactions, and the complexity of the	350
		query, ensuring both responsiveness and legal rele-	351
		vance.	352

Model	Year: 2022				Year: 2023			
	W/o RAG	Naive RAG	2-Step RAG	Tools	W/o RAG	Naive RAG	2-Step RAG	Tools
<i>Random Choice</i>	25.00	25.20	25.40	25.20	25.60	25.40	25.20	25.00
<i>Proprietary Models</i>								
GPT-4o	18.60	62.40	69.20	73.60	19.20	58.80	67.80	72.20
Gemini-2.0-Flash	12.40	61.20	68.60	69.20	12.20	59.40	69.00	70.20
Gemini-2.5-Flash	<b>30.20</b>	<b>68.80</b>	<b>75.60</b>	<b>77.00</b>	<b>32.40</b>	<b>69.20</b>	<b>76.40</b>	<b>77.00</b>
<i>Open-Source Models: Small</i>								
Llama3.2-1B-Instruct	6.20	6.00	6.40	8.20	7.00	7.20	7.40	9.20
Llama3.2-3B-Instruct	9.20	9.00	11.20	13.40	11.20	11.40	14.00	15.20
Gemma-3-4B-it	12.40	15.20	20.60	23.20	14.40	16.60	22.20	26.40
Qwen3-4B-Instruct-2507	<b>14.20</b>	<b>28.40</b>	44.60	<b>49.80</b>	<b>16.20</b>	<b>29.40</b>	45.40	<b>52.40</b>
Command-A-8B	8.20	25.20	<b>47.00</b>	47.40	11.20	23.40	<b>49.20</b>	50.20
Llama3.1-8B-Instruct	15.20	18.40	30.20	32.40	16.20	18.60	33.40	35.20
<i>Open-Source Models: Large</i>								
Gemma-3-12B-it	20.40	35.20	48.40	50.20	22.60	36.20	52.40	54.20
Gemma-3-27B-it	28.20	45.40	60.20	64.40	30.20	46.60	62.40	66.20
Qwen3-30B-A3B-Instruct-2507	<b>34.20</b>	<b>50.40</b>	<b>65.60</b>	<b>70.80</b>	<b>36.20</b>	<b>52.40</b>	<b>67.20</b>	<b>72.40</b>
Llama3.1-70B-Instruct	24.20	30.40	40.20	42.40	26.20	32.20	44.40	46.20

Table 1: Evaluation of our legal assistant agent MINA on BD Bar Council Exam: MCQ. Scores are 5-run averages. Best in each category and setup are marked with **bold underline**.

## 4 Experiments and Evaluation

### 4.1 Evaluation Setup

To evaluate the system’s real-world utility as a legal reasoning agent, we benchmarked its performance against the Bangladesh Bar Council Examinations from 2022 and 2023. These exams represent a high-stakes national standard for entry into the legal profession and are conducted in three stages: *Multiple Choice Questions (MCQ)*, *Written*, and *Viva Voce*. Our evaluation pipeline was designed to mirror this full progression. Each exam instance was administered under conditions closely aligned with actual testing protocols. For the MCQ and written stages, the model was assessed using original exam questions. For MCQ, each setup and model combination is repeated 5 times and average is taken to reduce model variability. In the case of written and viva evaluations, we partnered with law faculty members and legal experts from leading Bangladeshi universities (names anonymized) who served as independent evaluators. Each written answer was assessed by at least two qualified judges, and the final score was determined by averaging the two, following Bar Council standards.

### 4.2 Evaluation Metrics

We simulate the real-world evaluation process of the Bangladesh Bar Council examination across three major components: Multiple Choice Questions (MCQs), Written Examination, and Viva Voce. This design ensures that model performance can be meaningfully compared to that of human

candidates under authentic assessment conditions.

#### Multiple Choice Questions (MCQs):

For the MCQ component, automatic marking was performed by comparing each model’s responses to the ground truth, replicating the Optical Mark Recognition (OMR)-based evaluation used in actual examinations. Scores were computed on a 100-point scale, where a minimum of 50% is required for human candidates to qualify for the written stage according to the official Bar Council standard.

**Written Exam:** The written examination was evaluated through a human assessment framework aligned with the official marking criteria of the Bangladesh Bar Council. Performance was judged across four primary dimensions: accuracy, clarity, contextual understanding, and legal reasoning. Each evaluator applied a standardized rubric but provided a single composite score rather than separate ratings for each dimension, reflecting the holistic marking approach used in real examinations. The written section consisted of thirteen questions. The first eleven were organized into five sets, from which six questions were required to be answered—two from the first set and one from each of the remaining four—constituting ninety marks in total. The remaining two questions, forming Set F, were valued at ten marks, with candidates required to answer only one. For evaluation, the questions were selected based on the model’s confidence and performance consistency to ensure representative and fair benchmarking.

**Viva Voce:** The viva voce was conducted through

Model	Year: 2022				Year: 2023			
	W/o RAG	Naive RAG	2-Step RAG	Tools	W/o RAG	Naive RAG	2-Step RAG	Tools
<b>Proprietary Models</b>								
GPT-4o	19.20	55.40	69.80	71.20	21.20	60.80	72.60	75.60
Gemini-2.0-Flash	18.40	58.20	68.60	70.00	18.20	65.40	74.80	75.20
Gemini-2.5-Flash	<b>35.20</b>	<b>70.40</b>	<b>78.60</b>	<b>81.00</b>	<b>36.20</b>	<b>71.20</b>	<b>79.40</b>	<b>81.80</b>
<b>Open-Source Models: Small</b>								
Llama3.2-1B-Instruct	6.20	6.00	6.20	7.00	7.00	7.00	7.20	8.00
Llama3.2-3B-Instruct	10.20	11.00	14.20	16.40	12.00	14.20	17.40	20.00
Gemma-3-4B-it	15.20	20.20	24.40	28.00	18.20	22.40	26.00	30.20
Qwen3-4B-Instruct-2507	22.20	56.40	<b>71.60</b>	68.00	26.20	60.40	64.20	70.40
Command-A-8B	25.00	<b>60.20</b>	71.00	<b>74.40</b>	25.20	<b>73.00</b>	<b>74.20</b>	<b>76.00</b>
Llama3.1-8B-Instruct	<b>28.20</b>	34.40	46.20	50.00	<b>30.20</b>	66.00	68.20	71.20
<b>Open-Source Models: Large</b>								
Gemma-3-12B-it	30.00	42.20	55.40	58.00	32.20	44.40	58.00	60.20
Gemma-3-27B-it	38.20	55.00	68.20	72.40	40.00	56.20	70.00	74.20
Qwen3-30B-A3B-Instruct-2507	42.20	60.40	74.00	78.20	44.00	62.20	76.40	79.40
Llama3.1-70B-Instruct	<b>45.00</b>	<b>62.20</b>	<b>75.00</b>	<b>79.80</b>	<b>47.20</b>	<b>65.00</b>	<b>77.20</b>	<b>80.20</b>

Table 2: Evaluation of our legal assistant agent MINA on BD Bar Council Exam: Written. Scores are 5-evaluator averages. Best in each color-category and setup are marked with **bold underline**.

Model	W/o RAG	N. RAG	2-S RAG	Tools
<b>Proprietary Models</b>				
Gemini-2.0 Flash	32.80	60.80	74.40	76.60
Gemini-2.5 Flash	<b>36.20</b>	<b>70.40</b>	<b>79.20</b>	<b>81.00</b>
<b>Open-Source Models: Small</b>				
Llama3.2-1B-Instruct	6.20	6.40	6.20	7.00
Llama3.2-3B-Instruct	11.20	12.40	15.20	17.40
Gemma-3-4B-it	16.20	50.40	65.20	69.20
Qwen3-4B-Instruct-2507	22.20	<b>56.40</b>	67.20	70.20
Command A	<b>27.80</b>	55.60	<b>70.40</b>	<b>71.20</b>
Llama3.1-8B-Instruct	24.20	52.20	66.40	67.20
<b>Open-Source Models: Large</b>				
Gemma-3-12B-it	31.20	43.20	56.40	59.20
Gemma-3-27B-it	39.20	56.00	69.20	72.40
Qwen3-30B-A3B-Instruct	42.20	61.20	75.20	79.40
Llama3.1-70B-Instruct	<b>46.00</b>	<b>63.20</b>	<b>77.20</b>	<b>80.20</b>

Table 3: Evaluation of Viva exam across different setups. Scores are 5-evaluator averages. Best in each category and setup are marked with **bold underline**.

a ChatGPT-like interactive platform that simulated the real-world oral examination setting. The same evaluative principles applied as in the written component, with a minimum average of 50% required for human candidates to pass this stage. Evaluators assessed the model’s ability to sustain coherent, contextually appropriate, and legally accurate dialogue through natural, conversational exchanges resembling human viva sessions. Although the official Bar Council outcome for this stage is recorded simply as “pass” or “fail,” evaluators in our study assigned detailed scores out of 100 to provide granular feedback and enable comparative analysis. All scoring rubrics and thresholds were aligned with historical Bar Council evaluation criteria to ensure fairness, consistency, and comparability with human performance.

More details are available in Appendix F.

### 4.3 Baselines and Human Performance

To contextualize model performance, we compared results with actual candidate statistics from the 2022 and 2023 Bangladesh Bar Council Exams.

**Multiple Choice Questions (MCQs):** In 2022, only 25.86% of candidates (10,527 out of 40,696) passed the MCQ stage. In 2023, the pass rate declined further to 17.96% (6,229 out of 34,682). Although the passing threshold is set at just 50%, these low success rates highlight the complexity and difficulty of the questions, which are often challenging to interpret and answer correctly.

**Written Exam:** The written stage includes 13 questions across six legal domains, with candidates required to answer seven in total. Human examinees typically achieve 40–60% in this stage. In 2022, 53.94% of candidates (5,533 out of 10,527) passed; in 2023, this figure declined to 44.21% (2,754 out of 6,229). Our model consistently outperformed these benchmarks, demonstrating generalization and adaptability across exam years.

**Viva Voce:** The final stage evaluates legal articulation, judgment, and argumentative coherence. Pass rates for this stage are notably high, as the viva is essentially an oral extension of the written exam, and candidates who reach this point are typically well-prepared. In 2022, 96.65% of candidates (5,348 out of 5,533) passed the viva, while in 2023, 97.25% (2,973 out of 3,057; some participants were conditionally allowed) successfully cleared this stage.

The evaluation of MINA was conducted by law faculty from leading Bangladeshi universities, with five evaluators scoring viva voce responses and

three scoring written exams. Evaluators applied official Bar Council criteria for accuracy, clarity, contextual understanding, and legal reasoning. Inter-annotator agreement was high (Cohen’s Kappa = 0.827); numerical disagreements were resolved by averaging scores, and qualitative feedback was aggregated to identify recurring error patterns.

## 5 Results and Findings

We evaluated system performance in four setups: without RAG, one stage RAG (without act based filtering), two step RAG, and with all tools.

**Performance on Preliminary MCQ.** Results in Table 1 reveal some patterns in model performance on the Bangladesh Bar Council MCQ exam. Closed-source proprietary models (Gemini family) maintain a clear advantage, with Gemini-2.5-Flash consistently outperforming all baselines across years and setups, particularly in the multi-step RAG + tools condition where it exceeds 77%. Within open-source models, scale and architecture are decisive: smaller models like Llama3.2-1B and Gemma-3-4B-it barely surpass random-choice baselines, whereas large-scale Qwen3-30B achieves competitive performance, approaching proprietary levels under retrieval-augmented settings. Retrieval consistently improves results: naive RAG yields moderate gains, while 2-step RAG with tool integration substantially boosts performance, especially for large open-source models. Qwen3-30B shows steeper improvements than Gemma-27B or Llama-70B, highlighting the importance of data alignment and retrieval synergy for legal reasoning. Performance gaps between 2022 and 2023 remain small, indicating stability across cohorts rather than year-specific overfitting. These findings underscore both the promise and limitations of open-source LLMs in specialized legal reasoning: scaling helps, but architecture and training data choices remain critical.

**Performance on Written Exam.** Written evaluation results in Table 2 reveal both parallels and divergences from the MCQ setting. Proprietary models dominate, with Gemini-2.5-Flash achieving the highest scores across years, consistently surpassing 80% with tools. Open-source models demonstrate greater competitiveness in written tasks compared to MCQs: large-scale models like Llama3.1-70B and Qwen3-30B reach 79–81% with multi-step RAG and tools. Smaller models, such as Command-A-8B and Qwen3-4B, show significant gains when retrieval is added, highlighting

the benefit of structured augmentation. Year-to-year consistency indicates generalization across cohorts rather than memorization. Scaling advantages are more pronounced in written evaluation, where moving to 27B or 70B substantially boosts performance. These results suggest that open-source models, paired with robust RAG pipelines, can provide cost-effective, competitive legal reasoning, especially in Global South contexts like Bangladesh.

**Performance on Viva Voce (conducted via Chat).** Viva Voce evaluation (Table 3) shows clear trends across model classes and RAG setups. Proprietary models, especially Gemini-2.5 Flash, achieve top scores across all configurations, reaching 81.0 with tools. Among open-source models, scale correlates strongly with performance: small models like Llama3.2-1B score minimally, whereas Qwen3-30B and Llama3.1-70B exceed 75 with advanced RAG and tools. Retrieval significantly boosts performance, especially for smaller and mid-sized models, with naive or 2-step RAG increasing scores by 20–40 points. Tool integration further provides consistent, marginal gains for large models. Overall, large open-source models with retrieval and reasoning pipelines can approach proprietary performance, while small models remain underpowered, highlighting the importance of scale, domain-specific retrieval, and tool-assisted reasoning.

**Comparison with Human Examinee Performance.** Compared to human performance, the models demonstrate remarkable competitiveness across all three stages. In MCQs, Gemini-2.5 and Qwen3-30B consistently exceed the human pass threshold of 50%, despite candidate pass rates as low as 17.96% in 2023. In written exams, large open-source models reach 79–81%, matching or slightly exceeding typical human averages (40–60%). In the viva, optimized models score 70–76%, below human near-certainty (>96%) but still exhibiting legally coherent reasoning. These results suggest that LLMs can surpass most candidates in knowledge-intensive stages and approach human-like standards in oral reasoning, raising policy-relevant questions about AI-augmented professional legal assessment and the potential for AI to approach elite human performance, particularly in text-based evaluations. Detailed analysis is available in Appendix I.

**Cost Analysis.** Cost analysis (§B) shows that per-query inference remains extremely low across model families: MCQ-style queries cost roughly 0.2–0.6 cents and longer written queries 0.8–2.0

cents, with Qwen and LLaMA consistently cheaper than Gemini, and non-LLM tool overheads negligible. Even under a conservative multi-turn upper bound of 10 cents ( $\approx$  D12.2 BDT), a full interaction represents only about 0.12%–0.61% of the typical 2,000–10,000 BDT cost of basic legal consultation in Bangladesh. These results demonstrate that MINA can deliver jurisdiction-specific legal assistance at a price point compatible with large-scale, public access-to-justice deployment.

## 6 Error Analysis

**MCQ.** As detailed in Appendix G.1, systematic errors reveal that the model often misinterprets Bengali conjunctions (e.g., “O” (and) vs. “ba”(or)), flattening nuanced legal semantics. Jurisdictional hierarchy errors indicate weak integration of procedural layering with lexical parsing, while doctrinal misclassifications (e.g., *Res Judicata* under Section 151) expose failures in mapping abstract concepts to codified provisions. Intra-order confusions (Order 1, Rules 8 vs. 13) further suggest reliance on surface cues over legal reasoning. These patterns show the need for structured legal knowledge, hierarchical reasoning, and fine-grained disambiguation to enhance procedural accuracy.

**Written.** As discussed in Appendix G.2, the written-response analysis reveals recurring failure modes that reflect both knowledge limitations and reasoning weaknesses. The model occasionally employs imprecise legal terminology, omits critical conditions required for complete answers, and inconsistently distinguishes between closely related legal domains. Errors also arise from shallow multi-step reasoning, contextual misinterpretation, and a tendency toward fluent but unsupported generation, which can reduce factual reliability. These patterns indicate that while the model exhibits strong structural and linguistic competence, it remains vulnerable in high-stakes settings without explicit reasoning constraints and verification mechanisms. Addressing these issues requires improved domain-adaptive training, structured reasoning support, and validation against authoritative legal sources.

## 7 Evaluator Response Analysis

**Written.** As summarized in Appendix H.1, evaluators found that retrieval-augmented models generated coherent, exam-style responses following the IRAC structure, with strong statutory recall and drafting comparable to human scripts. Recurring weaknesses included omitted procedural

prerequisites, arithmetic errors in limitation calculations, and incomplete cross-statute synthesis. While bilingual clarity and structured reasoning were praised, evaluators noted inconsistent statutory precision and occasional hallucinated citations. Overall, the results reflect exam-ready fluency but highlight the need for deterministic procedural calculators, statutory-fusion mechanisms, and fine-tuning on annotated bilingual exam data to reach distinction-level accuracy.

**Viva.** As detailed in Appendix H.2, viva evaluations showed that MINA maintained composure, clear diction, and accurate statutory articulation under conversational pressure. Examiners valued its coherence and explanatory clarity but observed limited adaptability to factual variations and a lack of clarifying questions, sometimes resulting in incorrect procedural guidance. Rapid exchanges revealed occasional section misattributions treated as major oral faults. Evaluators emphasized that distinction-level performance requires greater interpretive flexibility, uncertainty awareness, and policy reasoning, motivating system-level enhancements such as clarification policies, dialogue-state tracking, and real-time procedural validation.

## 8 Concluding Remarks

We present MINA, a multilingual LLM-based legal assistant for the Bangladeshi legal system, designed to improve access to justice for low-income and linguistically diverse users through two-stage RAG, multilingual embeddings, and selective tool augmentation. Across MCQ, written, and viva-style evaluations, mid-sized open-source models match average human performance, while stronger models approach ceiling-level results, all within a modular, lightweight architecture suitable for resource-constrained deployment and enhanced by dictionary-based bilingual processing of colonial-era terminology. Cost analysis shows that using the AI agent yields an approximate 99.4%–99.9% reduction relative to traditional human-provided legal services, enabling economically sustainable, large-scale deployment. Error analysis reveals persistent challenges in procedural hierarchy, temporal reasoning, and adaptive oral dialogue, demonstrating that retrieval alone is insufficient and that targeted fine-tuning, structured procedural modules, real-time verification, and clarification policies are essential for reliable high-stakes legal assistance in low-resource, multilingual settings.

670	<b>Limitations</b>		
671	While our evaluation demonstrates notable performance gains through retrieval and generation strategies, several limitations remain. Retrieval quality is highly dependent on the underlying corpus; noisy or misaligned documents can still mislead even robust pipelines. Although strategies like Two Step RAG improve performance, they introduce additional latency and complexity that may not scale well in real-time systems.		
680	<b>Ethical Considerations</b>		
681	We adhered to all ethical guidelines outlined by the Association for Computational Linguistics (ACL) throughout this study. All data used in our benchmark were either publicly available or ethically sourced with appropriate permissions where required. No personally identifiable information (PII) was collected, stored, or used in the experiments. Human annotators involved in dataset construction were fairly compensated and provided with clear task instructions. We ensured transparency in our evaluation pipeline and made efforts to avoid biased or culturally insensitive content. Furthermore, model outputs were reviewed to identify and mitigate potential ethical harms. All experiments were conducted in accordance with ACL’s code of ethics regarding research integrity, fairness, and respect for contributors.		
698	<b>Potential Risks</b>		
699	While MINA is designed as a supportive tool for legal professionals, legal aid workers, and exam preparation, it is not infallible, and incorrect outputs remain possible, particularly when prompts are ambiguous, incomplete, or involve highly specialized legal scenarios. Erroneous advice, if relied upon without verification, could lead to misinformed decisions, procedural errors, or unintended legal consequences, especially in high-stakes cases. Liability for any legal action remains with the human professional using the system, consistent with standard practices for legal research platforms and drafting tools. There is also a risk of over-reliance, where users might defer critical judgment to the AI, potentially reducing diligence or critical thinking. To mitigate these risks, MINA should always be used under human supervision, and outputs must be cross-checked against authoritative statutes, case law, and professional guidance. On the positive		
	side, when properly integrated, the system can enhance lawyer productivity, reduce research time, improve accessibility for underserved populations, and support more equitable legal services by lowering operational costs and enabling broader dissemination of legal knowledge. Overall, careful deployment with oversight and clear guidelines is essential to maximize benefits while minimizing potential harms.		718 719 720 721 722 723 724 725 726
	<b>Reproducibility Statement</b>		727
	To ensure full reproducibility of our experiments, all code, trained retrieval indices, and processed datasets will be made publicly available upon acceptance. Detailed instructions for setting up the environment, running the Orchestrator Agent, executing the two-stage RAG pipeline, and reproducing all evaluations for MCQs, written exams, and viva voce are included in the paper and repository. Additionally, prompts, scoring rubrics, and evaluation protocols are documented to allow independent verification of both quantitative results and qualitative analyses.		728 729 730 731 732 733 734 735 736 737 738 739
	<b>References</b>		740
	Arif Ahmed. 2021. Challenges with access to justice in bangladesh. <i>Centre for Governance Studies</i> .		741 742
	Farzana Akter. 2017. <a href="#">Legal aid for ensuring access to justice in bangladesh: A paradox?</a> <i>Asian Journal of Law and Society</i> , 4(1):257–275.		743 744 745
	Farid Ariai and Gianluca Demartini. 2025. <a href="#">Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges</a> . <i>Preprint</i> , arXiv:2410.21306.		746 747 748 749
	Tanim Asjad. 2023. <a href="#">Use of bangla in judicial system</a> . Accessed: 2025-06-22.		750 751
	Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammam, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry		752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767



891	Akrass, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal,	955
892	Praveen Krishnan, Punit Singh Koura, Puxin Xu,	956
893	Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj	957
894	Ganapathy, Ramon Calderer, Ricardo Silveira Cabral,	958
895	Robert Stojnic, Roberta Raileanu, Rohan Maheswari,	959
896	Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron-	960
897	nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan	961
898	Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa-	962
899	hana Chennabasappa, Sanjay Singh, Sean Bell, Seo-	963
900	hyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sha-	964
901	ran Narang, Sharath Raparthy, Sheng Shen, Shengye	965
902	Wan, Shruti Bhosale, Shun Zhang, Simon Van-	966
903	denhende, Soumya Batra, Spencer Whitman, Sten	967
904	Sootla, Stephane Collot, Suchin Gururangan, Syd-	968
905	ney Borodinsky, Tamar Herman, Tara Fowler, Tarek	969
906	Sheasha, Thomas Georgiou, Thomas Scialom, Tobias	970
907	Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal	971
908	Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh	972
909	Ramanathan, Viktor Kerkez, Vincent Conguet, Vir-	973
910	ginie Do, Vish Vogeti, Vitor Albiero, Vladan Petro-	974
911	vic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit-	975
912	ney Meers, Xavier Martinet, Xiaodong Wang, Xi-	976
913	aofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xin-	977
914	feng Xie, Xuchao Jia, Xuwei Wang, Yaelle Gold-	978
915	schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen,	979
916	Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao,	980
917	Zacharie Delpierre Coudert, Zheng Yan, Zhengxing	981
918	Chen, Zoe Papakipos, Aaditya Singh, Aayushi Sri-	982
919	vastava, Abha Jain, Adam Kelsey, Adam Shajnfeld,	983
920	Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand,	984
921	Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei	985
922	Baevski, Allie Feinstein, Amanda Kallet, Amit San-	986
923	gani, Amos Teo, Anam Yunus, Andrei Lupu, An-	987
924	dres Alvarado, Andrew Caples, Andrew Gu, Andrew	988
925	Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchan-	989
926	dani, Annie Dong, Annie Franco, Anuj Goyal, Apar-	990
927	ajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,	991
928	Ashwin Bharambe, Assaf Eisenman, Azadeh Yaz-	992
929	dan, Beau James, Ben Maurer, Benjamin Leonhardi,	993
930	Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi	994
931	Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock,	995
932	Bram Wasti, Brandon Spence, Brani Stojkovic,	996
933	Brian Gamido, Britt Montalvo, Carl Parker, Carly	997
934	Burton, Catalina Mejia, Ce Liu, Changan Wang,	998
935	Changkyu Kim, Chao Zhou, Chester Hu, Ching-	999
936	Hsiang Chu, Chris Cai, Chris Tindal, Christoph Fe-	1000
937	ichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty,	1001
938	Daniel Kreymer, Daniel Li, David Adkins, David	1002
939	Xu, Davide Testuggine, Delia David, Devi Parikh,	1003
940	Diana Liskovich, Didem Foss, Dingkan Wang, Duc	1004
941	Le, Dustin Holland, Edward Dowling, Eissa Jamil,	1005
942	Elaine Montgomery, Eleonora Presani, Emily Hahn,	1006
943	Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban	1007
944	Arcaute, Evan Dunbar, Evan Smothers, Fei Sun,	1008
945	Felix Kreuk, Feng Tian, Filippus Kokkinos, Firat	1009
946	Ozgenel, Francesco Caggioni, Frank Kanayet, Frank	1010
947	Seide, Gabriela Medina Florez, Gabriella Schwarz,	1011
948	Gada Badeer, Georgia Swee, Gil Halpern, Grant	1012
949	Herman, Grigory Sizov, Guangyi, Zhang, Guna	1013
950	Lakshminarayanan, Hakan Inan, Hamid Shojanaz-	1014
951	eri, Han Zou, Hannah Wang, Hanwen Zha, Haroun	1015
952	Habeeb, Harrison Rudolph, Helen Suk, Henry Aspe-	1016
953	gren, Hunter Goldman, Hongyuan Zhan, Ibrahim	1017
954	Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,	1018
	Irina-Elena Veliche, Itai Gat, Jake Weissman, James	
	Geboski, James Kohli, Janice Lam, Japhet Asher,	
	Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-	
	nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy	
	Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe	
	Cummings, Jon Carvill, Jon Shepard, Jonathan Mc-	
	Phie, Jonathan Torres, Josh Ginsburg, Junjie Wang,	
	Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khan-	
	delwal, Katayoun Zand, Kathy Matosich, Kaushik	
	Veeraraghavan, Kelly Michelena, Keqian Li, Ki-	
	ran Jagadeesh, Kun Huang, Kunal Chawla, Kyle	
	Huang, Lailin Chen, Lakshya Garg, Lavender A,	
	Leandro Silva, Lee Bell, Lei Zhang, Liangpeng	
	Guo, Licheng Yu, Liron Moshkovich, Luca Wehrst-	
	edt, Madian Khabsa, Manav Avalani, Manish Bhatt,	
	Martynas Mankus, Matan Hasson, Matthew Lennie,	
	Matthias Reso, Maxim Groshev, Maxim Naumov,	
	Maya Lathi, Meghan Keneally, Miao Liu, Michael L.	
	Seltzer, Michal Valko, Michelle Restrepo, Mihir Pa-	
	tel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark,	
	Mike Macey, Mike Wang, Miquel Jubert Hermoso,	
	Mo Metanat, Mohammad Rastegari, Munish Bansal,	
	Nandhini Santhanam, Natascha Parks, Natasha	
	White, Navyata Bawa, Nayan Singhal, Nick Egebo,	
	Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich	
	Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz,	
	Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin	
	Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pe-	
	dro Rittner, Philip Bontrager, Pierre Roux, Piotr	
	Dollar, Polina Zvyagina, Prashant Ratanchandani,	
	Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel	
	Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu	
	Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,	
	Raymond Li, Rebekkah Hogan, Robin Battey, Rocky	
	Wang, Russ Howes, Ruty Rinott, Sachin Mehta,	
	Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara	
	Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov,	
	Satadru Pan, Saurabh Mahajan, Saurabh Verma,	
	Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-	
	say, Shaun Lindsay, Sheng Feng, Shenghao Lin,	
	Shengxin Cindy Zha, Shishir Patil, Shiva Shankar,	
	Shuqiang Zhang, Shuqiang Zhang, Sinong Wang,	
	Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala,	
	Stephanie Max, Stephen Chen, Steve Kehoe, Steve	
	Satterfield, Sudarshan Govindaprasad, Sumit Gupta,	
	Summer Deng, Sungmin Cho, Sunny Virk, Suraj	
	Subramanian, Sy Choudhury, Sydney Goldman, Tal	
	Remez, Tamar Glaser, Tamara Best, Thilo Koehler,	
	Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim	
	Matthews, Timothy Chou, Tzook Shaked, Varun	
	Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai	
	Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad	
	Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu,	
	Vladimir Ivanov, Wei Li, Wenchen Wang, Wen-	
	wen Jiang, Wes Bouaziz, Will Constable, Xiaocheng	
	Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo	
	Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia,	
	Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,	
	Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao,	
	Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary	
	DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang,	
	Zhiwei Zhao, and Zhiyu Ma. 2024. <a href="#">The llama 3 herd</a>	
	<a href="#">of models</a> . <i>Preprint</i> , arXiv:2407.21783.	

1019	Péter Homoki and Zsolt Zódi. 2024. <a href="#">Large language models and their possible uses in law</a> . <i>Hungarian Journal of Legal Studies</i> , 64(3):435–455.	1074
1020		1075
1021		1076
1022	Jia-Hong Huang, Chao-Chun Yang, Yixian Shen, Alessio M. Paccès, and Evangelos Kanoulas. 2024. <a href="#">Optimizing numerical estimation and operational efficiency in the legal domain through large language models</a> . <i>Preprint</i> , arXiv:2407.19041.	1077
1023		1078
1024		1079
1025		1080
1026		1081
1027	Atin Sakkeer Hussain and Anu Thomas. 2024. <a href="#">Large language models for judicial entity extraction: A comparative study</a> . <i>Preprint</i> , arXiv:2407.05786.	1082
1028		1083
1029		
1030	Md Jahirul Islam, Masahiro Suzuki, and Nurunnahar Mazumder. 2024. <a href="#">Promoting access to justice in bangladesh: Towards a hybrid justice model</a> . <i>International Journal of Law, Crime and Justice</i> , 77:100655.	1084
1031		1085
1032		1086
1033		1087
1034	Arafat Reza Jaan. 2023. <a href="#">What makes access to justice elusive in bangladesh?</a> Accessed: 2025-06-22.	1088
1035		1089
1036	Cong Jiang and Xiaolei Yang. 2024. <a href="#">Agents on the bench: Large language model based multi agent framework for trustworthy digital justice</a> . <i>Preprint</i> , arXiv:2412.18697.	1090
1037		1091
1038		1092
1039		1093
1040	Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. <a href="#">Leveraging large language models for learning complex legal concepts through storytelling</a> . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 7194–7219, Bangkok, Thailand. Association for Computational Linguistics.	1094
1041		1095
1042		1096
1043		1097
1044		1098
1045		1099
1046		1100
1047		1101
1048		1102
1049	Md Ibrahim Khalilullah. 2025. <a href="#">Justice on hold: Addressing bangladesh’s judicial backlog</a> . Accessed: 2025-06-23.	1103
1050		1104
1051		1105
1052	Shammem Ara Knock, Dr. Md. Omar Faruk, Rukhsana Siddiqua, and Asma Akter. 2024. <a href="#">The practice of rule of law in criminal justice system: A criminological study of bangladesh</a> . <i>International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences</i> , 12(4).	1106
1053		1107
1054		1108
1055		1109
1056		1110
1057		1111
1058	Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023. <a href="#">Large language models in law: A survey</a> . <i>Preprint</i> , arXiv:2312.03718.	1112
1059		1113
1060		1114
1061	Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2024. <a href="#">Legalagentbench: Evaluating llm agents in legal domain</a> . <i>Preprint</i> , arXiv:2412.17259.	1115
1062		1116
1063		1117
1064		1118
1065		1119
1066	Guannan Liang and Qianqian Tong. 2025. <a href="#">Llm-powered ai agent systems and their applications in industry</a> . <i>Preprint</i> , arXiv:2505.16120.	1120
1067		1121
1068		1122
1069	Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. 2024. <a href="#">Interpretable long-form legal question answering with retrieval-augmented large language models</a> . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 38(20):22266–22275.	1123
1070		1124
1071		1125
1072		1126
1073		1127
		1128
		1129
		1130
		1131
		1132
		1133
	Marios Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, and Konstantinos Tarabanis. 2024a. <a href="#">A large language model based legal assistant for governance applications</a> .	1074
		1075
		1076
		1077
	Marios Evangelos Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, and Konstantinos Tarabanis. 2024b. <a href="#">A Large Language Model Agent Based Legal Assistant for Governance Applications</a> , page 286–301. Springer Nature Switzerland.	1078
		1079
		1080
		1081
		1082
		1083
	Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. 2024. <a href="#">Better call gpt, comparing large language models against lawyers</a> . <i>Preprint</i> , arXiv:2401.16212.	1084
		1085
		1086
		1087
	Eliza Mik. 2024. <a href="#">Caveat lector: Large language models in legal practice</a> . <i>arXiv preprint</i> .	1088
		1089
	Arpeeta Shams Mizan. 2021. <a href="#">The law of language and the language of the law: A sociolegal appraisal of colonial legal language in bangladesh</a> . <i>Dhaka University Law Journal</i> , 32(2):119–134.	1090
		1091
		1092
		1093
	OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang,	1094
		1095
		1096
		1097
		1098
		1099
		1100
		1101
		1102
		1103
		1104
		1105
		1106
		1107
		1108
		1109
		1110
		1111
		1112
		1113
		1114
		1115
		1116
		1117
		1118
		1119
		1120
		1121
		1122
		1123
		1124
		1125
		1126
		1127
		1128
		1129
		1130
		1131
		1132
		1133

1134	Fred von Lohmann, Freddie Sulit, Gabriel Goh,	Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan	1198
1135	Gene Oden, Geoff Salmon, Giulio Starace, Greg	Cheu, Saachi Jain, Sam Altman, Sam Schoenholz,	1199
1136	Brockman, Hadi Salman, Haiming Bao, Haitang	Sam Toizer, Samuel Miserendino, Sandhini Agar-	1200
1137	Hu, Hannah Wong, Haoyu Wang, Heather Schmidt,	wal, Sara Culver, Scott Ethersmith, Scott Gray, Sean	1201
1138	Heather Whitney, Heewoo Jun, Hendrik Kirchner,	Grove, Sean Metzger, Shamez Hermani, Shantanu	1202
1139	Henrique Ponde de Oliveira Pinto, Hongyu Ren,	Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shi-	1203
1140	Huiwen Chang, Hyung Won Chung, Ian Kivlichan,	rong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay,	1204
1141	Ian O’Connell, Ian O’Connell, Ian Osband, Ian Sil-	Srinivas Narayanan, Steve Coffey, Steve Lee, Stew-	1205
1142	ber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya	art Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao	1206
1143	Kostrikov, Ilya Sutskever, Ingmar Kanitscheider,	Xu, Tarun Gogineni, Taya Christianson, Ted Sanders,	1207
1144	Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub	Tejal Patwardhan, Thomas Cunninghamman, Thomas	1208
1145	Pachocki, James Aung, James Betker, James Crooks,	Degry, Thomas Dimson, Thomas Raoux, Thomas	1209
1146	James Lennon, Jamie Kiros, Jan Leike, Jane Park,	Shadwell, Tianhao Zheng, Todd Underwood, Todor	1210
1147	Jason Kwon, Jason Phang, Jason Teplitz, Jason	Markov, Toki Sherbakov, Tom Rubin, Tom Stasi,	1211
1148	Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Var-	Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce	1212
1149	avva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui	Walters, Tyna Eloundou, Valerie Qi, Veit Moeller,	1213
1150	Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang,	Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne	1214
1151	Joaquin Quinonero Candela, Joe Beutler, Joe Lan-	Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra,	1215
1152	ders, Joel Parish, Johannes Heidecke, John Schul-	Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian,	1216
1153	man, Jonathan Lachman, Jonathan McKay, Jonathan	Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen	1217
1154	Uesato, Jonathan Ward, Jong Wook Kim, Joost	He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and	1218
1155	Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross,	Yury Malkov. 2024. <i>Gpt-4o system card. Preprint,</i>	1219
1156	Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao,	arXiv:2410.21276.	1220
1157	Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai	Aidan Peppin, Julia Kreutzer, Alice Schoenauer Se-	1221
1158	Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin	bag, Kelly Marchisio, Beyza Ermis, John Dang,	1222
1159	Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu,	Samuel Cahyawijaya, Shivalika Singh, Seraphina	1223
1160	Kenny Nguyen, Keren Gu-Lemberg, Kevin Button,	Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-	1224
1161	Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle	Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh	1225
1162	Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lau-	Fadaee, and Sara Hooker. 2025. <i>The multilingual</i>	1226
1163	ren Workman, Leher Pathak, Leo Chen, Li Jing, Lia	<i>divide and its impact on global ai safety. Preprint,</i>	1227
1164	Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lil-	arXiv:2505.21344.	1228
1165	lian Weng, Lindsay McCallum, Lindsey Held, Long	Weicong Qin and Zhongxiang Sun. 2024. <i>Exploring the</i>	1229
1166	Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kon-	<i>nexus of large language models and legal systems: A</i>	1230
1167	draciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz,	<i>short survey. Preprint, arXiv:2404.00990.</i>	1231
1168	Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine	Md. Alamgir Sarkar Raj. 2024. <i>Assessing the legal</i>	1232
1169	Boyd, Madeleine Thompson, Marat Dukhan, Mark	<i>framework and socioeconomic impacts of compensa-</i>	1233
1170	Chen, Mark Gray, Mark Hudnall, Marvin Zhang,	<i>tion for wrongfully convicted and imprisoned persons</i>	1234
1171	Marwan Aljubei, Mateusz Litwin, Matthew Zeng,	<i>in bangladesh: Challenges and policy recommenda-</i>	1235
1172	Max Johnson, Maya Shetty, Mayank Gupta, Meghan	<i>tions. Cognizance Journal of Multidisciplinary Stud-</i>	1236
1173	Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao	<i>ies, 4(6):456–466.</i>	1237
1174	Zhong, Mia Glaese, Mianna Chen, Michael Jan-	Laura Safdie. 2025. <i>Ai and legal aid: A generational</i>	1238
1175	ner, Michael Lampe, Michael Petrov, Michael Wu,	<i>opportunity for access to justice. Accessed: 2025-06-</i>	1239
1176	Michele Wang, Michelle Fradin, Michelle Pokrass,	<i>23.</i>	1240
1177	Miguel Castro, Miguel Oom Temudo de Castro,	Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo	1241
1178	Mikhail Pavlov, Miles Brundage, Miles Wang, Minal	Chang, and Yang Li. 2024. <i>Lawluo: A chi-</i>	1242
1179	Khan, Mira Murati, Mo Bavarian, Molly Lin,	<i>nese law firm co-run by llm agents. Preprint,</i>	1243
1180	Murat Yesildal, Nacho Soto, Natalia Gimelshein, Na-	arXiv:2407.16252.	1244
1181	talie Cone, Natalie Staudacher, Natalie Summers,	Ummey Sharaban Tahura and Shawkat Alam. 2025. <i>Le-</i>	1245
1182	Natan LaFontaine, Neil Chowdhury, Nick Ryder,	<i>gal aid, systemic barriers, and access to justice in</i>	1246
1183	Nick Stathas, Nick Turley, Nik Tezak, Niko Felix,	<i>bangladesh. Accessed: 2025-06-23.</i>	1247
1184	Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel	Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-	1248
1185	Bundick, Nora Puckett, Ofir Nachum, Ola Okelola,	Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan	1249
1186	Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins,	Schalkwyk, Andrew M. Dai, Anja Hauth, Katie	1250
1187	Olivier Godement, Owen Campbell-Moore, Patrick	Millican, David Silver, Melvin Johnson, Ioannis	1251
1188	Chao, Paul McMillan, Pavel Belov, Peng Su, Peter	Antonoglou, Julian Schrittwieser, Amelia Glaese,	1252
1189	Bak, Peter Bakkum, Peter Deng, Peter Dolan,	Jilin Chen, Emily Pitler, Timothy Lillicrap, Ange-	1253
1190	Peter Hoeschele, Peter Welinder, Phil Tillet, Philip	liki Lazaridou, Orhan Firat, James Molloy, Michael	1254
1191	Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming	Isard, Paul R. Barham, Tom Hennigan, Benjamin	1255
1192	Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Ra-		
1193	jan Troll, Randall Lin, Rapha Gontijo Lopes, Raul		
1194	Puri, Reah Miyara, Reimar Leike, Renaud Gaubert,		
1195	Reza Zamani, Ricky Wang, Rob Donnelly, Rob		
1196	Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchan-		
1197	dani, Romain Huet, Rory Carmichael, Rowan Zellers,		

1256	Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong	Zilka, Flavien Prost, Luheng He, Marianne Mon-	1319
1257	Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza	teiro, Gaurav Mishra, Chris Welty, Josh Newlan,	1320
1258	Rutherford, Erica Moreira, Kareem Ayoub, Megha	Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu,	1321
1259	Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-	Raoul de Liedekerke, Justin Gilmer, Carl Saroufim,	1322
1260	Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty	Shruti Rijhwani, Shaobo Hou, Disha Shrivastava,	1323
1261	Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao	Anirudh Baddepudi, Alex Goldin, Adnan Ozturel,	1324
1262	Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah,	Albin Cassirer, Yunhan Xu, Daniel Sohn, Deven-	1325
1263	Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran,	dra Sachan, Reinald Kim Amplayo, Craig Swan-	1326
1264	Sumit Bagri, Balaji Lakshminarayanan, Jeremiah	son, Dessie Petrova, Shashi Narayan, Arthur Guez,	1327
1265	Liu, Andras Orban, Fabian Güra, Hao Zhou, Xiny-	Siddhartha Brahma, Jessica Landon, Miteyan Pa-	1328
1266	ing Song, Aurelien Boffy, Harish Ganapathy, Steven	tel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wen-	1329
1267	Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu,	hao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung,	1330
1268	Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej	James Keeling, Petko Georgiev, Diana Mincu, Boxi	1331
1269	Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa,	Wu, Salem Haykal, Rachel Saputro, Kiran Vodra-	1332
1270	Majd Al Merey, Martin Baeuml, Zhifeng Chen, Lau-	halli, James Qin, Zeynep Cankara, Abhanshu Sharma,	1333
1271	rent El Shafey, Yujing Zhang, Olcan Sercinoglu,	Nick Fernando, Will Hawkins, Behnam Neyshabur,	1334
1272	George Tucker, Enrique Piqueras, Maxim Krikun,	Solomon Kim, Adrian Hutter, Priyanka Agrawal,	1335
1273	Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca	Alex Castro-Ros, George van den Driessche, Tao	1336
1274	Roelofs, Anaïs White, Anders Andreassen, Tamara	Wang, Fan Yang, Shuo yiin Chang, Paul Komarek,	1337
1275	von Glehn, Lakshman Yagati, Mehran Kazemi, Lu-	Ross McIlroy, Mario Lučić, Guodong Zhang, Wael	1338
1276	cas Gonzalez, Misha Khalman, Jakob Sygnowski,	Farhan, Michael Sharman, Paul Natsev, Paul Michel,	1339
1277	Alexandre Frechette, Charlotte Smith, Laura Culp,	Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shak-	1340
1278	Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan	eri, Christina Butterfield, Justin Chung, Paul Kishan	1341
1279	Schucher, Federico Lebron, Alban Rrustemi, Na-	Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar	1342
1280	talie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao,	Soparkar, Karel Lenc, Timothy Chung, Aedan Pope,	1343
1281	Bartek Perz, Dian Yu, Heidi Howard, Adam Blo-	Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo	1344
1282	niarz, Jack W. Rae, Han Lu, Laurent Sifre, Mar-	Wang, Joshua Maynez, Mary Phuong, Taylor Tobin,	1345
1283	cello Maggioni, Fred Alcober, Dan Garrette, Megan	Andrea Tacchetti, Maja Trebacz, Kevin Robinson,	1346
1284	Barnes, Shantanu Thakoor, Jacob Austin, Gabriel	Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan	1347
1285	Barth-Maron, William Wong, Rishabh Joshi, Rahma	Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone,	1348
1286	Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh	Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gri-	1349
1287	Tomar, Evan Senter, Martin Chadwick, Ilya Kor-	bovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music	1350
1288	nakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu,	Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers,	1351
1289	Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia,	Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed,	1352
1290	Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse	Tianqi Liu, Richard Powell, Vijay Bolina, Mariko	1353
1291	Hartman, Xavier Garcia, Thanumalayan Sankara-	Inuma, Polina Zablotskaia, James Besley, Da-Woon	1354
1292	narayana Pillai, Jacob Devlin, Michael Laskin, Diego	Chung, Timothy Dozat, Ramona Comanescu, Xi-	1355
1293	de Las Casas, Dasha Valter, Connie Tao, Lorenzo	ance Si, Jeremy Greer, Guolong Su, Martin Polacek,	1356
1294	Blanco, Adrià Puigdomènech Badia, David Reitter,	Raphaël Lopez Kaufman, Simon Tokumine, Hexiang	1357
1295	Mianna Chen, Jenny Brennan, Clara Rivera, Sergey	Hu, Elena Buchatskaya, Yingjie Miao, Mohamed	1358
1296	Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski,	Elhawaty, Aditya Siddhant, Nenad Tomasev, Jin-	1359
1297	Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yim-	wei Xing, Christina Greer, Helen Miller, Shereen	1360
1298	ing Gu, Kate Olszewska, Ravi Addanki, Antoine	Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Ange-	1361
1299	Miech, Annie Louis, Denis Teplyashin, Geoff Brown,	los Filos, Milos Besta, Rory Blevins, Ted Klimenko,	1362
1300	Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang,	Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Os-	1363
1301	Zoe Ashwood, Anton Briukhov, Albert Webson, San-	car Chang, Mantas Pajarskas, Carrie Muir, Vered	1364
1302	jay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-	Cohen, Charline Le Lan, Krishna Haridasan, Amit	1365
1303	Wei Chang, Axel Stjerngren, Josip Djolonga, Yut-	Marathe, Steven Hansen, Sholto Douglas, Rajku-	1366
1304	ing Sun, Ankur Bapna, Matthew Aitchison, Pedram	mar Samuel, Mingqiu Wang, Sophia Austin, Chang	1367
1305	Pejman, Henryk Michalewski, Tianhe Yu, Cindy	Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo,	1368
1306	Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich,	Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gle-	1369
1307	Kehang Han, Peter Humphreys, Thibault Sellam,	icher, Thi Avrahami, Anudhyan Boral, Hansa Srini-	1370
1308	James Bradbury, Varun Godbole, Sina Samangooui,	vasan, Vittorio Selo, Rhys May, Konstantinos Aiso-	1371
1309	Bogdan Damoc, Alex Kaskasoli, Sébastien M. R.	pos, Léonard Hussenot, Livio Baldini Soares, Kate	1372
1310	Arnold, Vijay Vasudevan, Shubham Agrawal, Jason	Baumli, Michael B. Chang, Adrià Recasens, Ben	1373
1311	Riesa, Dmitry Lepikhin, Richard Tanburn, Srivat-	Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo,	1374
1312	san Srinivasan, Hyeontaek Lim, Sarah Hodgkinson,	Anita Gergely, Justin Frye, Vinay Ramasesh, Dan	1375
1313	Pranav Shyam, Johan Ferret, Steven Hand, Ankush	Horgan, Kartikeya Badola, Nora Kassner, Subhra-	1376
1314	Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Gi-	jit Roy, Ethan Dyer, Víctor Campos Campos, Alex	1377
1315	ang, Alexander Neitz, Zaheer Abbas, Sarah York,	Tomala, Yunhao Tang, Dalia El Badawy, Elspeth	1378
1316	Machel Reid, Elizabeth Cole, Aakanksha Chowdh-	White, Basil Mustafa, Oran Lang, Abhishek Jin-	1379
1317	ery, Dipanjan Das, Dominika Rogozińska, Vitaliy	dal, Sharad Vikram, Zhitao Gong, Sergi Caelles,	1380
1318	Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas	Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng,	1381
		Wojciech Stokowiec, Ce Zheng, Phoebe Thacker,	1382

1383	Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh,	Geoff Bacon, David Greene, Daniil Mirylenka, Chen	1447
1384	James Svensson, Max Bileschi, Piyush Patil, Ankesh	Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel	1448
1385	Anand, Roman Ring, Katerina Tsihlias, Arpi Vezzer,	Andermatt, Patrick Siegler, Ben Horn, Assaf Is-	1449
1386	Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom	rael, Francesco Pongetti, Chih-Wei "Louis" Chen,	1450
1387	Kwiatkowski, Samira Daruki, Keran Rong, Allan	Marco Selvatici, Pedro Silva, Kathie Wang, Jack-	1451
1388	Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg,	son Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai,	1452
1389	Mina Khan, Lisa Anne Hendricks, Marie Pellat,	Alessandro Agostini, Maulik Shah, Hung Nguyen,	1453
1390	Vladimir Feinberg, James Cobon-Kerr, Tara Sainath,	Noah Ó Donnaile, Sébastien Pereira, Linda Friso,	1454
1391	Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives,	Adam Stambler, Adam Kurzrok, Chenkai Kuang,	1455
1392	Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd,	Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang,	1456
1393	Le Hou, Qingze Wang, Thibault Sottiaux, Michela	Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qi-	1457
1394	Paganini, Jean-Baptiste Lespiau, Alexandre Mou-	jun Tan, Dan Banica, Daniel Balle, Ryan Pham,	1458
1395	farek, Samer Hassan, Kaushik Shivakumar, Joost van	Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot	1459
1396	Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh	Singh, Chris Hidey, Niharika Ahuja, Pranab Sax-	1460
1397	Goyal, Matthew Tung, Andrew Brock, Hannah Sheah-	ena, Dan Dooley, Srividya Pranavi Potharaju, Eileen	1461
1398	han, Vedant Misra, Cheng Li, Nemanja Rakićević,	O'Neill, Anand Gokulchandran, Ryan Foley, Kai	1462
1399	Mostafa Dehghani, Fangyu Liu, Sid Mittal, Jun-	Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta,	1463
1400	hyuk Oh, Seb Noury, Eren Sezener, Fantine Huot,	Ragha Kotikalapudi, Chalence Safranek-Shrader, An-	1464
1401	Matthew Lamm, Nicola De Cao, Charlie Chen, Sid-	drew Goodman, Joshua Kessinger, Eran Globen, Pra-	1465
1402	harth Mudgal, Romina Stella, Kevin Brooks, Gau-	teek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang	1466
1403	tam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita	Song, Ali Eichenbaum, Thomas Brovelli, Sahitya	1467
1404	Melinkeri, Aaron Cohen, Venus Wang, Kristie Sey-	Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani,	1468
1405	more, Sergey Zubkov, Rahul Goel, Summer Yue,	Charles Chen, Andy Crawford, Shalini Pal, Mukund	1469
1406	Sai Krishnakumaran, Brian Albert, Nate Hurley,	Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski,	1470
1407	Motoki Sano, Anhad Mohanane, Jonah Joughin,	Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Nic-	1471
1408	Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiaw-	colò Dal Santo, Siddharth Goyal, Jitesh Punjabi,	1472
1409	ern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor	Karthik Kappaganthu, Chester Kwak, Pallavi LV,	1473
1410	Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara	Sarmishta Velury, Himadri Choudhury, Jamie Hall,	1474
1411	Padmanabhan, Subha Puttagunta, Kalpesh Krishna,	Premal Shah, Ricardo Figueira, Matt Thomas, Min-	1475
1412	Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam	jie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi,	1476
1413	Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin,	Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak,	1477
1414	Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Si-	Victor Åhdel, Sujeevan Rajayogam, Travis Choma,	1478
1415	ciliano, Alan Papir, Robby Neale, Jonas Bragagnolo,	Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vin-	1479
1416	Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang,	cent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie	1480
1417	Richie Feng, Milad Gholami, Kevin Ling, Lijuan	Zhou, Mehrdad Khatir, Charles Sutton, Wojciech	1481
1418	Liu, Jules Walter, Hamid Moghaddam, Arun Kishore,	Rzadkowski, Fiona Macintosh, Roopali Vij, Kon-	1482
1419	Jakub Adamek, Tyler Mercado, Jonathan Mallinson,	stantin Shagin, Paul Medina, Chen Liang, Jinjing	1483
1420	Siddhinita Wandekar, Stephen Cagle, Eran Ofek,	Zhou, Pararth Shah, Yingying Bi, Attila Dankovics,	1484
1421	Guillermo Garrido, Clemens Lombriser, Maksim	Shipra Banga, Sabine Lehmann, Marissa Bredesen,	1485
1422	Mukha, Botu Sun, Hafeezul Rahman Mohammad,	Zifan Lin, John Eric Hoffmann, Jonathan Lai, Ray-	1486
1423	Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus,	nald Chung, Kai Yang, Nihal Balani, Arthur Braziń-	1487
1424	Quan Yuan, Leif Schelin, Oana David, Ankur Garg,	skas, Andrei Sozanschi, Matthew Hayes, Héctor Fer-	1488
1425	Yifan He, Oleksii Duzhyi, Anton Ålmyr, Timo-	nández Alcalde, Peter Makarov, Will Chen, Anto-	1489
1426	thée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex	nio Stella, Liselotte Snijders, Michael Mandl, Ante	1490
1427	Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie	Kärroman, Pawel Nowak, Xinyi Wu, Alex Dyck, Kr-	1491
1428	Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed,	ishnan Vaidyanathan, Raghavender R, Jessica Mal-	1492
1429	Subhabrata Das, Zihang Dai, Kyle He, Daniel von	let, Mitch Rudominer, Eric Johnston, Sushil Mit-	1493
1430	Dincklage, Shyam Upadhyay, Akanksha Maurya,	tal, Akhil Udatum, Janara Christensen, Vishal Verma,	1494
1431	Luyan Chi, Sebastian Krause, Khalid Salama, Pam G	Zach Irving, Andreas Santucci, Gamaleldin Elsayed,	1495
1432	Rabinovitch, Pavan Kumar Reddy M, Aarush Sel-	Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan	1496
1433	van, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Gu-	Hua, Geoffrey Cideron, Edouard Leurent, Mah-	1497
1434	ven, Himanshu Gupta, Boyi Liu, Deepak Sharma,	moud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy	1498
1435	Idan Heimlich Shtacher, Shachi Paul, Oscar Aker-	Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper	1499
1436	lund, François-Xavier Aubet, Terry Huang, Chen	Shnok, Mukund Sundararajan, Xuezhi Wang, Zack	1500
1437	Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze,	Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar,	1501
1438	Francesco Bertolini, Liana-Eleonora Marinescu, Mar-	Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan	1502
1439	tin Bülle, Dominik Paulus, Khyatti Gupta, Tejasi	Uesato, Romina Datta, Oskar Bunyan, Shimu Wu,	1503
1440	Latkar, Max Chang, Jason Sanders, Roopa Wil-	John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner,	1504
1441	son, Xuwei Wu, Yi-Xuan Tan, Lam Nguyen Thiet,	Subhajit Naskar, Michael Azzam, Matthew Johnson,	1505
1442	Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming	Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez	1506
1443	Chen, Thang Luong, Seth Benjamin, Jasmine Lee,	Elias, Afroz Mohiuddin, Faizan Muhammad, Jin	1507
1444	Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan,	Miao, Andrew Lee, Nino Vieillard, Jane Park, Ji-	1508
1445	Krzysztof Styrz, Pengcheng Yin, Jon Simon, Mal-	ageng Zhang, Jeff Stanway, Drew Garmon, Abhijit	1509
1446	colm Rose Harriott, Mudit Bansal, Alexei Robsky,		

1510	Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Lu-	Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck,	1574
1511	owei Zhou, Jonathan Evens, William Isaac, Geoffrey	Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoff-	1575
1512	Irving, Edward Loper, Michael Fink, Isha Arkatkar,	mann, Dan Holtmann-Rice, Olivier Bachem, Sho	1576
1513	Nanxin Chen, Izhak Shafran, Ivan Petrychenko,	Arora, Christy Koh, Soheil Hassas Yeganeh, Siim	1577
1514	Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai	Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita,	1578
1515	Zhu, Peter Grabowski, Yu Mao, Alberto Magni,	Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, An-	1579
1516	Kaisheng Yao, Javier Snaider, Norman Casagrande,	mol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz,	1580
1517	Evan Palmer, Paul Suganthan, Alfonso Castaño,	Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown,	1581
1518	Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński,	Shreya Singh, Wei Fan, Aaron Parisi, Joe Stan-	1582
1519	Ashwin Sreevatsa, Jennifer Prendki, David Soergel,	ton, Vinod Koverkathu, Christopher A. Choquette-	1583
1520	Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari,	Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash	1584
1521	Meenu Gaba, Jeremy Wiesner, Diana Gage Wright,	Shroff, Mani Varadarajan, Sanaz Bahargam, Rob	1585
1522	Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay	Willoughby, David Gaddy, Guillaume Desjardins,	1586
1523	Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu,	Marco Cornero, Brona Robenek, Bhavishya Mit-	1587
1524	Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert	tal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev,	1588
1525	Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith	Henrik Jacobsson, Alireza Ghaffarkhah, Morgane	1589
1526	Pallo, Abhishek Chakladar, Ginger Perng, Elena Al-	Rivière, Alanna Walton, Clément Crepy, Alicia Par-	1590
1527	lica Abellan, Mingyang Zhang, Ishita Dasgupta,	rish, Zongwei Zhou, Clement Farabet, Carey Rade-	1591
1528	Nate Kushman, Ivo Penchev, Alena Repina, Xihui	baugh, Praveen Srinivasan, Claudia van der Salm,	1592
1529	Wu, Tom van der Weide, Priya Ponnappalli, Car-	Andreas Fidjeland, Salvatore Scellato, Eri Latorre-	1593
1530	oline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier	Chimoto, Hanna Klimczak-Plucińska, David Bridson,	1594
1531	Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pa-	Dario de Cesare, Tom Hudson, Piermaria Mendolic-	1595
1532	sumarathi, Nathan Lintz, Anitha Vijayakumar, Daniel	chio, Lexi Walker, Alex Morris, Matthew Mauger,	1596
1533	Andor, Pedro Valenzuela, Minnie Lui, Cosmin Padu-	Alexey Guseynov, Alison Reid, Seth Odoom, Lu-	1597
1534	raru, Daiyi Peng, Katherine Lee, Shuyuan Zhang,	cia Loher, Victor Cotruta, Madhavi Yenugula, Do-	1598
1535	Somer Greene, Duc Dung Nguyen, Paula Kurylow-	minik Grewe, Anastasia Petrushkina, Tom Duerig,	1599
1536	icz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam	Antonio Sanchez, Steve Yadlowsky, Amy Shen,	1600
1537	Choo, Ziqiang Feng, Biao Zhang, Achintya Sing-	Amir Globerson, Lynette Webb, Sahil Dua, Dong	1601
1538	hal, Dayou Du, Dan McKinnon, Natasha Antropova,	Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi,	1602
1539	Tolga Bolukbasi, Orgad Keller, David Reid, Daniel	Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj	1603
1540	Finchelstein, Maria Abi Raad, Remi Crocker, Peter	Khare, Shreyas Rammohan Belle, Lei Wang, Chetan	1604
1541	Hawkins, Robert Dadashi, Colin Gaffney, Ken	Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin	1605
1542	Franko, Anna Bulanova, Rémi Leblond, Shirley	Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao	1606
1543	Chung, Harry Askham, Luis C. Cobo, Kelvin Xu,	Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Man-	1607
1544	Felix Fischer, Jun Xu, Christina Sorokin, Chris Al-	ish Reddy Vuyyuru, John Aslanides, Nidhi Vyas,	1608
1545	berti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev,	Martin Wicke, Xiao Ma, Evgenii Elytshev, Nina Mar-	1609
1546	Hannah Forbes, Dylan Banarse, Zora Tung, Mark	tin, Hardie Cate, James Manyika, Keyvan Amiri,	1610
1547	Omernick, Colton Bishop, Rachel Sterneck, Rohan	Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier,	1611
1548	Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno,	Nilesh Tripuraneni, David Madras, Mandy Guo,	1612
1549	Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz,	Austin Waters, Oliver Wang, Joshua Ainslie, Jason	1613
1550	Alex Polozov, Victoria Krakovna, Sasha Brown, Mo-	Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer,	1614
1551	hammadHossein Bateni, Dennis Duan, Vlad Firoiu,	Feng Yang, Riham Mansour, Jason Gelman, Yang Xu,	1615
1552	Meghana Thotakuri, Tom Natan, Matthieu Geist,	George Polovets, Ji Liu, Honglong Cai, Warren Chen,	1616
1553	Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko	XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof	1617
1554	Tojo, Michael Kwong, James Lee-Thorp, Christo-	Angermueller, Xiaowei Li, Anoop Sinha, Weiren	1618
1555	pher Yew, Danila Sinopalnikov, Sabela Ramos, John	Wang, Julia Wiesinger, Emmanouil Koukoumidis,	1619
1556	Mellor, Abhishek Sharma, Kathy Wu, David Miller,	Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark	1620
1557	Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jen-	Goldenson, Parashar Shah, MK Blake, Hongkun Yu,	1621
1558	nifer Beattie, Emily Caveness, Libin Bai, Julian	Anthony Urbanowicz, Jennimaria Palomaki, Chrisan-	1622
1559	Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi	tha Fernando, Ken Durden, Harsh Mehta, Nikola	1623
1560	Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng,	Momchev, Elahe Rahimtoroghi, Maria Georgaki,	1624
1561	Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh,	Amit Raul, Sebastian Ruder, Morgan Redshaw, Jin-	1625
1562	Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin,	hyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li,	1626
1563	Daniel Toyama, Evan Rosen, Sasan Tavakkol, Lint-	Blake Hechtman, Parker Schuh, Milad Nasr, Kieran	1627
1564	ing Xue, Chen Elkind, Oliver Woodman, John Car-	Milan, Vladimir Mikulik, Juliana Franco, Tim Green,	1628
1565	penter, George Papamakarios, Rupert Kemp, Sushant	Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea	1629
1566	Kafle, Tanya Grunina, Rishika Sinha, Alice Tal-	Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshi-	1630
1567	bert, Diane Wu, Denese Owusu-Afriyie, Cosmo	tij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang,	1631
1568	Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna	Ke Ye, Jean Michel Sarr, Melanie Moranski Preston,	1632
1569	Narayana, Jing Li, Saaber Fatehi, John Wieting,	Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta,	1633
1570	Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura	Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi	1634
1571	Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi	M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric	1635
1572	Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Re-	Chu, Xuanyi Dong, Amruta Muthal, Senaka Buth-	1636
1573	beca Santamaria-Fernandez, Sonam Goenka, Wenny	pitiya, Sarthak Jauhari, Nan Hua, Urvashi Khan-	1637

1638	delwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Sha-	Hussein Hazimeh, Ian Ballantyne, Idan Szpektor,	1701
1639	har Drath, Avigail Dabush, Nan-Jiang Jiang, Har-	Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe	1702
1640	shal Godhia, Uli Sachs, Anthony Chen, Yicheng	Stanton, John Wieting, Jonathan Lai, Jordi Orbay,	1703
1641	Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai,	Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jy-	1704
1642	James Wang, Chen Liang, Jenny Hamer, Chun-Sung	otinder Singh, Kat Black, Kathy Yu, Kevin Hui, Ki-	1705
1643	Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít	ran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella	1706
1644	Listík, Mathias Carlen, Jan van de Kerkhof, Marcin	Valentine, Marina Coelho, Marvin Ritter, Matt Hoff-	1707
1645	Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova,	man, Matthew Watson, Mayank Chaturvedi, Michael	1708
1646	Richard Stefanec, Vitaly Gatsko, Christoph Hirn-	Moynihan, Min Ma, Nabila Babar, Natasha Noy,	1709
1647	schall, Ashwin Sethi, Xingyu Federico Xu, Chetan	Nathan Byrd, Nick Roy, Nikola Momchev, Nilay	1710
1648	Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Ke-	Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil	1711
1649	shav Dhandhanian, Manish Katyal, Akshay Gupta,	Botarda, Paul Caron, Paul Kishan Rubenstein, Phil	1712
1650	Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan	Culliton, Philipp Schmid, Pier Giuseppe Sessa, Ping-	1713
1651	Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin	mei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shiv-	1714
1652	Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera	anna, Renjie Wu, Renke Pan, Reza Rokni, Rob	1715
1653	Filippova, Abhipso Ghosh, Ben Limonchik, Bhar-	Willoughby, Rohith Vallu, Ryan Mullins, Sammy	1716
1654	gava Urala, Chaitanya Krishna Lanka, Derik Clive,	Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal,	1717
1655	Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak,	Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhat-	1718
1656	Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal	nagar, Sindhu Raghuram Panyam, Sivan Eiger, Sus-	1719
1657	Majmundar, Michael Alverson, Michael Kucharski,	san Zhang, Tianqi Liu, Trevor Yacovone, Tyler	1720
1658	Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo	Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vin-	1721
1659	Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim,	cent Roseberry, Vlad Feinberg, Vlad Kolesnikov,	1722
1660	Swetha Sankar, Vineet Shah, Lakshmi Ramachan-	Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam	1723
1661	droni, Xiangkai Zeng, Ben Bariach, Laura Wei-	Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Vic-	1724
1662	dinger, Tu Vu, Alek Andreev, Antoine He, Kevin	tor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao,	1725
1663	Hui, Shelem Kashem, Amar Subramanya, Sissie	Kat Black, Nabila Babar, Jessica Lo, Erica Mor-	1726
1664	Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam	eira, Luiz Gustavo Martins, Omar Sanseviero, Lu-	1727
1665	Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu,	cas Gonzalez, Zach Gleicher, Tris Warkentin, Va-	1728
1666	Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2025a.	hab Mirrokni, Evan Senter, Eli Collins, Joelle Bar-	1729
1667	<a href="#">Gemini: A family of highly capable multimodal mod-</a>	ral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias,	1730
1668	<a href="#">els. Preprint, arXiv:2312.11805.</a>	D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer,	1731
1669	Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya	Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray	1732
1670	Pathak, Nino Vieillard, Ramona Merhej, Sarah Per-	Kavukcuoglu, Clement Farabet, Elena Buchatskaya,	1733
1671	rin, Tatiana Matejovicova, Alexandre Ramé, Morg-	Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin,	1734
1672	ane Rivière, Louis Rouillard, Thomas Mesnard, Ge-	Sebastian Borgeaud, Olivier Bachem, Armand Joulin,	1735
1673	offrey Cideron, Jean bastien Grill, Sabela Ramos,	Alek Andreev, Cassidy Hardin, Robert Dadashi, and	1736
1674	Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo	Léonard Hussenot. 2025b. <a href="#">Gemma 3 technical report.</a>	1737
1675	Penchev, Gaël Liu, Francesco Visin, Kathleen Ke-	<i>Preprint, arXiv:2503.19786.</i>	1738
1676	nealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin,	Azmine Toushik Wasi, Wahid Faisal, Mst Rafia Is-	1739
1677	Robert Busa-Fekete, Alex Feng, Noveen Sachdeva,	lam, and Mahathir Mohammad Bappy. 2024. <a href="#">Ex-</a>	1740
1678	Benjamin Coleman, Yi Gao, Basil Mustafa, Iain	<a href="#">ploring possibilities of ai-powered legal assistance</a>	1741
1679	Barr, Emilio Parisotto, David Tian, Matan Eyal,	<a href="#">in bangladesh through large language modeling.</a>	1742
1680	Colin Cherry, Jan-Thorsten Peter, Danila Sinopal-	<i>Preprint, arXiv:2410.17210.</i>	1743
1681	nikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran	Sabine Wehnert. 2023. <a href="#">Justifiable artificial intelligence:</a>	1744
1682	Kazemi, Dan Malkin, Ravin Kumar, David Vilar,	<a href="#">Engineering large language models for legal applica-</a>	1745
1683	Idan Brusilovsky, Jiaming Luo, Andreas Steiner,	<a href="#">tions. arXiv preprint.</a>	1746
1684	Abe Friesen, Abhanshu Sharma, Abheesht Sharma,	Qiong Yan. 2023. <a href="#">Legal challenges of artificial intel-</a>	1747
1685	Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa	<a href="#">ligence in the field of criminal defense. Lecture</a>	1748
1686	Saade, Alex Feng, Alexander Kolesnikov, Alexei	<a href="#">Notes in Education Psychology and Public Media,</a>	1749
1687	Bendebury, Alvin Abdagic, Amit Vadi, Andrés	30(1):167–175.	1750
1688	György, André Susano Pinto, Anil Das, Ankur	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,	1751
1689	Bapna, Antoine Miech, Antoine Yang, Antonia Pat-	Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,	1752
1690	erson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot,	Chengen Huang, Chenxu Lv, Chujie Zheng, Dayi-	1753
1691	Bo Wu, Bobak Shahriari, Bryce Petriani, Charlie	heng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge,	1754
1692	Chen, Charline Le Lan, Christopher A. Choquette-	Haoran Wei, Huan Lin, Jialong Tang, Jian Yang,	1755
1693	Choo, CJ Carey, Cormac Brick, Daniel Deutsch,	Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi	1756
1694	Danielle Eisenbud, Dee Cattle, Derek Cheng, Dim-	Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai	1757
1695	itris Pappas, Divyashree Shivakumar Sreepathi-	Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao	1758
1696	halli, Doug Reid, Dustin Tran, Dustin Zelle, Eric	Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang,	1759
1697	Noland, Erwin Huiuzenga, Eugene Kharitonov, Fred-	Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan	1760
1698	erick Liu, Gagik Amirkhanyan, Glenn Cameron,	Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao	1761
1699	Hadi Hashemi, Hanna Klimczak-Plucińska, Har-		
1700	man Singh, Harsh Mehta, Harshal Tushar Lehri,		

- 1762 Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xu-  
1763 ancheng Ren, Yang Fan, Yang Su, Yichang Zhang,  
1764 Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang,  
1765 Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zi-  
1766 han Qiu. 2025. [Qwen3 technical report](#). *Preprint*,  
1767 arXiv:2505.09388.
- 1768 Sabina Yesmin. 2025. [Alternative dispute resolution:  
1769 A strategic solution to bangladesh’s judicial backlog](#).  
1770 Accessed: 2025-06-23.
- 1771 Zheng-Xin Yong, Beyza Ermis, Marzieh Fadaee,  
1772 Stephen H. Bach, and Julia Kreutzer. 2025. [The  
1773 state of multilingual llm safety research: From mea-  
1774 suring the language gap to mitigating it](#). *Preprint*,  
1775 arXiv:2505.24119.
- 1776 Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang,  
1777 Shujun Liu, Yun Song, Xuanjing Huang, and  
1778 Zhongyu Wei. 2025. [Multi-agent simulator drives  
1779 language models for legal intensive interaction](#). In  
1780 *Findings of the Association for Computational Lin-  
1781 guistics: NAACL 2025*, pages 6537–6570, Albu-  
1782 querque, New Mexico. Association for Computa-  
1783 tional Linguistics.
- 1784 Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang  
1785 Zhang, Zhiyuan Liu, and Maosong Sun. 2020. [How  
1786 does NLP benefit legal system: A summary of legal  
1787 artificial intelligence](#). In *Proceedings of the 58th  
1788 Annual Meeting of the Association for Computational  
1789 Linguistics*, pages 5218–5230, Online. Association  
1790 for Computational Linguistics.

<b>A</b>	<b>Related Work</b>	<b>20</b>	<b>1791</b>
A.1	Legal NLP Research . . . . .	20	1792
A.2	Agent-focused Legal NLP Research . . . . .	20	1793
A.3	Multilingual Legal NLP Research . . . . .	20	1794
<b>B</b>	<b>Cost-per-Query Comparison and Affordability Implications</b>	<b>22</b>	<b>1795</b>
<b>C</b>	<b>Broader Impact</b>	<b>22</b>	<b>1796</b>
<b>D</b>	<b>System Demonstration</b>	<b>22</b>	<b>1797</b>
<b>E</b>	<b>Additional Tools for MINA</b>	<b>22</b>	<b>1798</b>
E.1	File Content Reader . . . . .	22	1799
E.2	Keyword Generator . . . . .	23	1800
E.3	Web Search and Web Page Parser . . . . .	23	1801
E.4	Question Relevance Analyzer . . . . .	24	1802
E.5	Chat Analyzer . . . . .	24	1803
E.6	Legal Dictionary . . . . .	24	1804
E.7	Socio-Economic Simulation . . . . .	24	1805
<b>F</b>	<b>Additional Information on Experiments and Evaluation</b>	<b>24</b>	<b>1806</b>
F.1	Evaluators . . . . .	24	1807
F.2	Evaluation Process of Written Exams . . . . .	25	1808
F.3	Inter-Annotator Agreement for Written Evaluation . . . . .	25	1809
F.4	Quality Assessment of Retrieved Sections . . . . .	25	1810
F.5	Real-World Evaluation . . . . .	26	1811
<b>G</b>	<b>Detailed Error Analysis</b>	<b>27</b>	<b>1812</b>
G.1	MCQ . . . . .	27	1813
G.2	Written Exam . . . . .	28	1814
G.2.1	Error Types . . . . .	31	1815
G.2.2	Root Causes . . . . .	31	1816
G.2.3	Potential Solutions . . . . .	32	1817
<b>H</b>	<b>Detailed Evaluator Response Analysis</b>	<b>32</b>	<b>1818</b>
H.1	Written Examination: detailed evaluation and patterns . . . . .	32	1819
H.2	Viva Voce (oral) Evaluation: dynamics, errors, and remediation . . . . .	34	1820
<b>I</b>	<b>Extended Analysis</b>	<b>35</b>	<b>1821</b>
I.1	Detailed Findings from MCQ . . . . .	35	1822
I.2	Detailed Findings from Written Exam . . . . .	36	1823
I.3	Detailed Findings from Viva Voce (Oral) . . . . .	37	1824
<b>J</b>	<b>More Ablation Studies</b>	<b>38</b>	<b>1825</b>
J.1	Tool Usage Analysis and Methodological Clarification . . . . .	38	1826
<b>K</b>	<b>Extended Discussion</b>	<b>39</b>	<b>1827</b>
<b>L</b>	<b>Prompts</b>	<b>41</b>	<b>1828</b>
L.1	Orchestrator Agent Prompt . . . . .	41	1829
L.2	User Prompt . . . . .	42	1830
L.3	RAG Agent Prompt . . . . .	43	1831

## A Related Work

### A.1 Legal NLP Research

Research in Legal Natural Language Processing and AI applications in law has advanced considerably, particularly for high-resource languages like English. Existing systems tackle diverse tasks such as legal document summarization, named entity recognition, question answering, contract analysis, and judgment prediction (Ariai and Demartini, 2025). Tools like ROSS Intelligence and Westlaw Edge automate research and e-discovery, while AI chatbots like DoNotPay assist with document drafting. Recent developments integrate large language models (LLMs) with logic programming and multi-agent frameworks leveraging knowledge graphs and Retrieval-Augmented Generation (RAG) to improve reliability and contextual accuracy. Furthermore, LLM-based methods employing specialized prompts and curated datasets for LegalAI tasks (Huang et al., 2024), as well as models like Mistral and Gemma for judicial entity extraction (Hussain and Thomas, 2024), demonstrate superior efficiency and cost-effectiveness over traditional practices (Sun et al., 2024). RAG and agent-based approaches enhance governance and long-form legal question answering (Mamalis et al., 2024a), while storytelling and prompt templates aid complex concept processing and text structuring (Jiang et al., 2024; de Kinderen and Winter, 2024). However, LLM limitations, such as stochastic outputs, hallucinations (Dahl et al., 2024), and fact-checking gaps (Mik, 2024) necessitate more interpretable, evidence-based models (Louis et al., 2024), especially for domain-sensitive contexts like law.

### A.2 Agent-focused Legal NLP Research

Prior work includes a GDPR-focused RAG and agent-based assistant using GPT-3.5/4 over EU legislation to answer complex queries, showing promising results with precision varying by query difficulty (Mamalis et al., 2024b). Chinese-language benchmarks like LegalAgentBench evaluate LLM agent performance across 17 corpora and tool-augmented workflows, offering nuanced metrics on reasoning and multi-hop tasks (Li et al., 2024). Multi-agent systems have also been explored: Chatlaw uses a MoE + knowledge graph approach to reduce hallucinations and mimic law-firm workflows (Cui et al., 2024), while frameworks like MASER and AgentsCourt simulate interactive legal scenarios and judicial deliberation using coordinated LLM-driven agents (Yue et al., 2025; Jiang and Yang, 2024). Broader surveys of LLM agents detail architectures integrating RAG, hierarchical planning, and safety mechanisms for domain-focused applications (Liang and Tong, 2025; Yong et al., 2025). Together, these efforts highlight a trend toward modular, tool-enhanced LLM agents tailored to legal reasoning, retrieval, and collaborative decision-making, forming the foundation for governance-focused assistants.

### A.3 Multilingual Legal NLP Research

Multilingual NLP research, crucial for low-resource languages, faces challenges from data scarcity, script variations, and cultural nuances (Yong et al., 2025; Peppin et al., 2025). Despite multilingual models like GPT-4 showing promise, performance degradation persists for underrepresented languages such as Bengali. Initiatives like MultiLegalPile (Qin and Sun, 2024) and legal text comprehension studies (Martin et al., 2024; Wehnert, 2023; Homoki and Zódi, 2024) aim to bridge this gap, though high-resource biases remain. Moreover, South Asian legal AI development is nascent, with limited empirical systems addressing jurisdiction-specific complexities. Tailoring LLMs to local law firms and regional practices (Homoki and Zódi, 2024) could enhance operational relevance. Wasi et al. (2024) explored enhancing Bangla capabilities of LLMs by fine-tuning GPT-2; however, the model still lacks robustness and comprehensive linguistic coverage.

Together, these pilot projects highlight AI’s potential to bridge justice gaps by automating legal research, document generation, and offering chatbot-based assistance. However, there remains a lack of deployed and empirically evaluated multilingual legal assistant systems that address the unique socio-legal and linguistic complexities of low resource but high population countries like Bangladesh. The proposed system seeks to fill this gap by providing a jurisdiction-specific, linguistically adapted, and socio-economically informed AI solution tailored to Bangladesh’s legal environment.

**🏠 Civil Suit Definition Error - "Property and Office" vs. "Property or Office"**

📄 Statement:  
 "দেওয়ানি প্রকৃতির মোকদ্দমা অর্থ এমন মোকদ্দমা যেখানে \_\_\_\_\_ থাকে। সংশ্লিষ্ট স্বার্থ জড়িত"

🗨️ Translation:  
 "A civil nature suit is one where \_\_\_\_\_ is involved. A related interest is at stake."

❌ Prediction: সম্পত্তি ও অফিস (*property and office*)

✅ Correct Answer: সম্পত্তি বা অফিস (*property or office*)

🔍 Analysis:  
 The key confusion lies in the conjunctive "ও" (**and**) vs. "বা" (**or**). Legally, a civil suit can concern **either** property **or** office—not necessarily both. The model mistakenly assumes **co-occurrence** where **alternation** is intended. This is a common linguistic trap in Bengali, where "ও" **adds specificity**, whereas "বা" **generalizes the scope**.

🧠 Linguistic Difficulty:  
 Low to Moderate. The legal meaning hinges on a simple **binary conjunction**, but its legal implication is precise—"বা" expands possible grounds for suits, while "ও" narrows it.

**🏠 Res Judicata Misclassification - Section 151 vs. Section 11**

📄 Statement:  
 "The Code of Civil Procedure, 1908 এর \_\_\_\_\_ ধারা একই বিচার্য বিষয় নিয়ে একই পক্ষগণের মধ্যে একাধিক বিচারকার্য নিষিদ্ধ করে।"

🗨️ Translation:  
 "Section \_\_\_\_\_ of the Code of Civil Procedure, 1908 prohibits multiple proceedings between the same parties on the same matter."

❌ Prediction: ১৫১ (Section 151)

✅ Correct Answer: ১১ (Section 11)

🔍 Analysis:  
 Section 11 legally codifies the doctrine of **Res Judicata**. Section 151 is a general **inherent powers clause**, not related to this doctrine. The model likely confused **technical phrasing like "multiple proceedings"** with general power-related provisions, indicating difficulty in **semantic linking of specialized legal doctrines to correct sections**.

🧠 Linguistic Difficulty:  
 High. Requires **mapping abstract legal doctrines** to codified sections, not just matching phrases. The term "একই বিচার্য বিষয়" (same matter) is a legal cue, but the model seems to prioritize numerically higher or more general sections.

**🏠 Jurisdictional Court Level Misunderstanding - "Any" vs. "Lowest" Court**

📄 Statement:  
 "প্রত্যেক মোকদ্দমা দাখিল করতে হবে বিচার এখতিয়ারসম্পন্ন \_\_\_\_\_ পর্যায়ে আদালত।"

🗨️ Translation:  
 "Each suit must be filed in a court having jurisdiction of the \_\_\_\_\_ level."

❌ Prediction: যে-কোনো (*any*)

✅ Correct Answer: সর্বনিম্ন (*lowest*)

🔍 Analysis:  
 This tests knowledge of the **principle of institution in the lowest competent court**, designed to reduce case backlog. The word "বিচার এখতিয়ারসম্পন্ন" (**competent court**) misled the model into thinking any court with jurisdiction suffices. However, legally, it must be the **lowest such court**.

🧠 Linguistic Difficulty:  
 Moderate. Misunderstanding arises from failing to combine "jurisdiction" + "hierarchy" conceptually. The model treats it as a **flat category**, not a layered system.

**🏠 Representative Suit Misidentification - Order 1, Rule 13 vs. Rule 8**

📄 Statement:  
 "The Code of Civil Procedure, 1908 এ 'প্রতিনিধিত্বমূলক মোকদ্দমা দায়েরের ধারণা পাওয়া যায় \_\_\_\_\_।"

🗨️ Translation:  
 "In the CPC, 1908, the concept of representative suit is found in \_\_\_\_\_."

❌ Prediction: Order 1, Rule 13

✅ Correct Answer: Order 1, Rule 8

🔍 Analysis:  
 Order 1, Rule 8 directly concerns **representative suits**, where many people with the same interest are represented collectively. Rule 13 concerns procedural objections. The model confuses **functionally distinct rules within the same Order**, indicating **difficulty with intra-order distinctions** in procedural law.

🧠 Linguistic Difficulty:  
 High. Needs precise **rule-level legal knowledge**, and the Bengali terms "প্রতিনিধিত্বমূলক মোকদ্দমা" (representative suit) don't appear in Rule 13, which may have misled the model.

Figure 2: Error Analysis (Command-A model examples)

## B Cost-per-Query Comparison and Affordability Implications

A central design requirement for access-to-justice systems in low-income settings is *economic sustainability*. Beyond accuracy and legal soundness, the feasibility of large-scale deployment is primarily determined by per-query inference cost. We therefore compare the approximate cost-per-query of representative proprietary and open-weight model families: *Gemini*, *Qwen*, and *LLaMA*, using empirical averages estimated from our prior experimental runs.

For short, structured MCQ-style queries, which typically require one to three tool calls and limited token generation, the estimated per-query cost lies in the range of 0.2–0.4 cents for Qwen (approximately 0.24–0.49 BDT), 0.3–0.5 cents for LLaMA (approximately 0.37–0.61 BDT), and 0.4–0.6 cents for Gemini (approximately 0.49–0.73 BDT). For longer written or constructed-response (CQ-type) queries, entailing two to four tool calls and substantially higher token usage, the corresponding costs increase to roughly 0.8–1.4 cents for Qwen (approximately 0.98–1.71 BDT), 1.0–1.6 cents for LLaMA (approximately 1.22–1.95 BDT), and 1.4–2.0 cents for Gemini (approximately 1.71–2.44 BDT). Across all configurations, the computational overhead of non-LLM tools, such as retrieval, translation, and formatting components, contributes negligibly to overall cost relative to LLM inference.

Even under a deliberately conservative upper-bound scenario involving a multi-turn interaction costing as much as 10 cents in total (approximately 12.2 BDT), the expense corresponds to only about 0.12%–0.61% of the minimum prevailing cost of basic legal advice in Bangladesh, which typically ranges from 2,000 to 10,000 BDT per consultation. Equivalently, using the AI agent represents a cost reduction of approximately 99.4%–99.9% relative to traditional human-provided legal services. When embedded within MINA’s multilingual, RAG-based chain-of-tools framework, these cost characteristics demonstrate that high-quality, jurisdiction-specific legal assistance can be delivered at a price point compatible with public-service deployment, reinforcing the practicality of scaling LLM-based systems to expand access to justice in low-resource settings.

## C Broader Impact

While the system was developed and evaluated within the context of Bangladesh’s legal framework, it is designed to be modular and adaptable, allowing potential deployment in other jurisdictions and low-resource languages. By replacing the legal corpus and updating the legal dictionary, the framework can be customized to different legal systems. Additionally, the use of multilingual embeddings facilitates scaling to languages beyond Bengali, including those with limited digital resources. This adaptability broadens the potential impact of the system, enabling cost-effective and contextually accurate AI-assisted legal reasoning in diverse global settings. Beyond the Bangladeshi context, such a framework could support legal education, case preparation, and access-to-justice initiatives in other low-resource or underrepresented jurisdictions, thereby promoting equitable access to legal knowledge and professional assistance worldwide.

## D System Demonstration

Figure 3 shows a demonstration of our system and UI.

## E Additional Tools for MINA

Our system integrates a suite of specialized tools that collectively enable structured, context-sensitive, and socially grounded legal reasoning. Each tool performs a distinct operational role within the overall workflow, from document parsing to contextual augmentation, ensuring that both procedural accuracy and interpretive sensitivity are maintained throughout the pipeline. The following subsections describe these tools in detail, outlining their design rationale, underlying mechanisms, and example applications.

### E.1 File Content Reader

The **File Content Reader** handles ingestion of uploaded materials across multiple formats, including *.pdf*, *.docx*, and *.pptx*. This module employs temporary file storage to maintain data privacy and

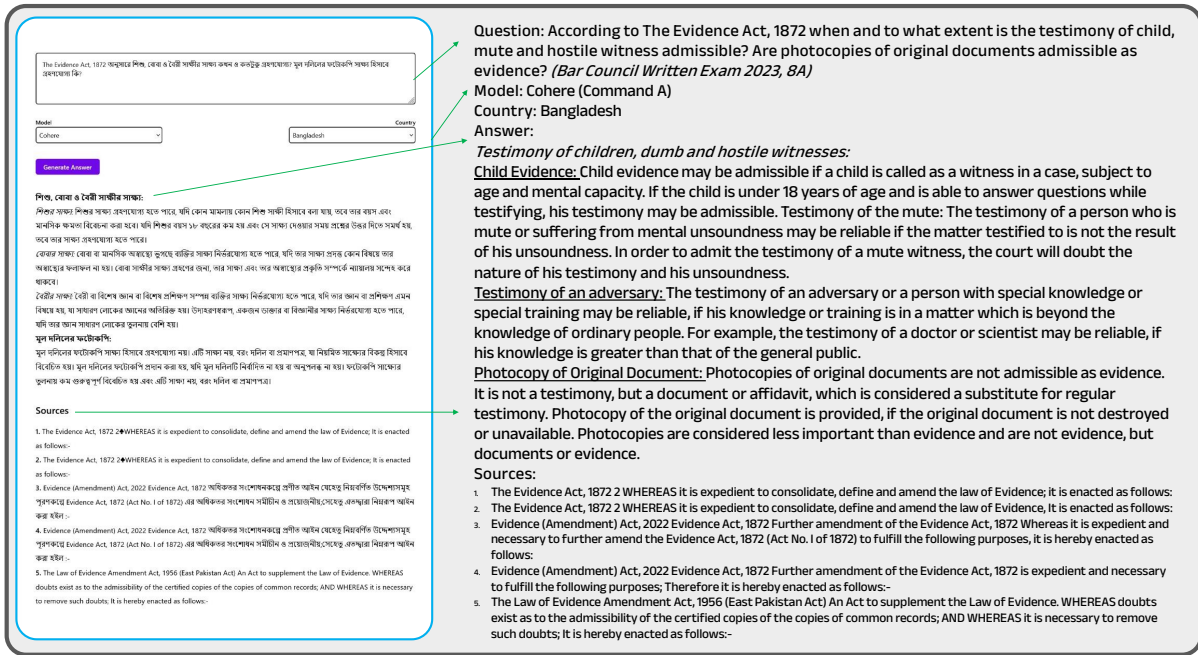


Figure 3: System Demonstration: UI and deployable system of MINA.

uses format-specific parsers such as `python-docx`, `pdfminer.six`, and `python-pptx` to extract clean, structured text. Non-textual artifacts (e.g., images, headers, footers, and embedded metadata) are automatically filtered to yield context-preserving, analysis-ready content.

Example: A user uploads a High Court Division judgment in `.pdf` format. The tool parses all pages, removes page numbers and footers, and produces a consolidated, section-labeled text block for semantic indexing. The processed text becomes directly usable for RAG-based retrieval or prompt conditioning, minimizing noise in subsequent reasoning steps.

## E.2 Keyword Generator

The **Keyword Generator** assists the retrieval-augmented generation (RAG) stage by producing a compact set of 5–10 semantically rich keywords derived from a user query or case prompt. It uses a lightweight LLM for semantic abstraction and includes a regular-expression-based fallback that ensures robust keyword extraction even under low-context or ambiguous input conditions.

Example: For the prompt “Draft a writ petition challenging unlawful termination under labor law,” the tool generates keyword clusters such as ["writ petition", "termination", "Bangladesh Labour Act", "fundamental rights", "Article 102", "judicial review"]. These keywords guide focused web retrieval and internal database searches, improving retrieval precision for domain-specific queries.

## E.3 Web Search and Web Page Parser

To integrate external knowledge dynamically, the **Web Search** module relies on DuckDuckGo’s query interface to fetch the top-ranked 3–5 search results, including page titles, URLs, and content snippets. When in-depth reading is required, the **Web Page Parser** uses BeautifulSoup to extract up to 5,000 characters of visible text while stripping scripts, navigation menus, and style elements.

Example: When tasked with “Summarize recent Supreme Court decisions on anticipatory bail in Bangladesh,” the modules work jointly to retrieve credible sources (e.g., *The Daily Star Law & Our Rights* or *Bangladesh Supreme Court Online*) and return excerpted paragraphs suitable for summarization or citation.

#### 1953 **E.4 Question Relevance Analyzer**

1954 The **Question Relevance Analyzer** maintains conceptual coherence in multi-turn dialogue by embedding  
1955 both current and previous queries into a shared semantic space. Cosine similarity between embeddings  
1956 determines whether the new query logically extends or diverges from the ongoing conversation.

1957 Example: When the user transitions from “Draft an appeal under Section 96 CPC” to “What  
1958 are the grounds for revision?”, the analyzer detects moderate semantic overlap but distinct  
1959 procedural context, prompting a structured reset rather than a contextual merge.

#### 1960 **E.5 Chat Analyzer**

1961 The **Chat Analyzer** provides an additional layer of context-awareness in conversational mode. It reviews  
1962 prior dialogue to reconstruct temporal dependencies, topic continuity, and user preferences, performing  
1963 discourse-level inference to recognize patterns such as clarification requests, corrections, or elaboration  
1964 prompts.

1965 Example: If a user first says “Draft a plaint for defamation” and later asks “Now make it fit  
1966 for filing before the Joint District Judge,” the Chat Analyzer detects continuity and refines the  
1967 procedural formatting automatically.

#### 1968 **E.6 Legal Dictionary**

1969 The **Legal Dictionary** module provides culturally adapted interpretive functionality. It explains colonial-  
1970 era, Farsi-influenced, or archaic terms still prevalent in Bangladeshi legal texts. Each term is annotated  
1971 with both its statutory definition and contemporary contextual meaning.

1972 Example: Input contains “*naraji*” (objection petition). The module outputs: “‘*Naraji*’ refers  
1973 to a formal objection against a police report, typically filed under Section 173(3) of the  
1974 CrPC, asserting dissatisfaction with the investigation outcome.” This facilitates bilingual legal  
1975 comprehension across English-Bengali legal ecosystems.

#### 1976 **E.7 Socio-Economic Simulation**

1977 The **Socio-Economic Simulation** module models how demographic and socio-economic variables in-  
1978 fluence access to justice, acknowledging that legal outcomes in Bangladesh often correlate with income,  
1979 occupation, literacy, and geography.

1980 Example: Consider Bar Council Exam 2023, Question 4(b) (translated): “X assaults Y. Enraged,  
1981 Y reacts violently. Taking advantage of Y’s anger, Z hands Y a knife intending Y will kill X.  
1982 On 03/11/2023, Y kills X at Rasulpur Bazaar. Police charges Y and Z under Sections 109, 34,  
1983 323, and 326 of the Penal Code, 1860. Prepare a ‘naraji’ petition before the court, determining  
1984 their criminal liability.”

1985 Simulation Input: - Accused Y: rural agricultural worker, monthly income 12,000 BDT, low  
1986 literacy - Accused Z: urban businessman - Victim X: marginalized community member

1987 Simulation Output: The tool predicts Y’s higher likelihood of prolonged pre-trial detention  
1988 due to inability to post bail, while Z is more likely to secure early representation. It highlights  
1989 systemic inequities, affordability, case delays, and social bias, that influence actual justice  
1990 outcomes beyond statutory law. This module transforms the system into a socio-legal diagnostic  
1991 platform rather than a purely doctrinal AI.

### 1992 **F Additional Information on Experiments and Evaluation**

#### 1993 **F.1 Evaluators**

1994 The evaluation of written and viva voce responses was conducted by law faculty members from leading  
1995 Bangladeshi universities. For the viva voce examination, five evaluators independently scored each  
1996 response, while three evaluators assessed the written examination. The evaluators applied the same

criteria used in official Bangladesh Bar Council examinations, including accuracy, clarity, contextual understanding, and legal reasoning. No additional training was provided, as all evaluators were already familiar with the official standards and marking rubrics.

Inter-annotator agreement was measured using Cohen’s Kappa, yielding a score of 0.827, indicating high consistency across evaluators. Disagreements in numerical scoring were resolved by averaging the scores of the five viva evaluators and three written evaluators. Qualitative feedback was aggregated to identify recurring error patterns, as detailed in the Error Analysis section.

This evaluation protocol ensures that the assessment closely mirrors real-world examination practices while maintaining reliability, fairness, and consistency across both written and oral components.

## F.2 Evaluation Process of Written Exams

The written evaluation of MINA was conducted by a panel of five law faculty members from leading Bangladeshi universities, each evaluating the system’s responses to all 13 questions of the 2022 and 2023 Bangladesh Bar Council Exams. Evaluators assessed answers along four dimensions: accuracy, clarity, contextual understanding, and legal reasoning. Each question was scored numerically (15 marks for Sets A–E, 10 marks for Set F), and the five scores were averaged to obtain a composite mark per question. The system generated answers for all questions, after which the model selected the “best answer” within each set with another prompt, reflecting realistic candidate strategy. Detailed qualitative feedback was provided for each answer, highlighting strengths such as logical structure, precise statutory references, and coherent reasoning. Weaknesses were noted in stepwise computation of limitation periods, explicit allocation of evidentiary burdens, and critical interpretation of procedural rules. Comparisons across 2022 and 2023 in Figure 4 showed minor year-to-year variations, with improvement in complex procedural questions such as injunctions and adjournments. Overall, the best-selected answers yielded a total score within 75–80% in good models, aligning with or surpassing average human candidate performance.

The evaluators’ feedback emphasized both doctrinal competence and practical procedural awareness (Appendix H.1). Comments highlighted the need to explicitly state preconditions, include stepwise calculations of limitation periods, and clarify evidentiary thresholds to achieve distinction-level responses. Some answers applied rules formulaically without sufficient critical analysis or discussion of judicial discretion. High-performing responses were praised for smooth argument transitions, clear mapping from statute to facts, and structured reasoning. Observed patterns indicate that MINA is strong in statutory interpretation and logical drafting but could benefit from enhancements in contextual sensitivity and explicit procedural reasoning. The selection of seven best answers for final scoring ensured realistic aggregation, reflecting typical candidate exam strategy. Overall, the evaluation demonstrates that a well-tuned LLM-based legal assistant can produce exam-quality written responses while revealing actionable areas for improvement in legal reasoning and clarity.

## F.3 Inter-Annotator Agreement for Written Evaluation

To quantify consistency among evaluators of the written exams, we calculated Cohen’s  $\kappa$  ( $\kappa$ ) between all pairs of evaluators. Each evaluator scored the 13 questions numerically along four dimensions: accuracy, clarity, contextual understanding, and legal reasoning. Cohen’s  $\kappa$  is computed as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where  $p_o$  is the observed proportion of agreement between evaluators and  $p_e$  is the expected agreement by chance. Across all questions and evaluator pairs, we obtained  $\kappa = 0.827$ , indicating high consistency. This demonstrates that, even with numerical scoring, our evaluation protocol yields reliable and reproducible assessments. Minor variations were observed in a few complex procedural questions, but overall agreement remained strong, supporting the robustness of our written exam evaluation methodology.

## F.4 Quality Assessment of Retrieved Sections

In our two-stage RAG pipeline, the relevance of retrieved legal sections is explicitly assessed before final response generation. An LLM-based filtering step evaluates each retrieved section against the user query

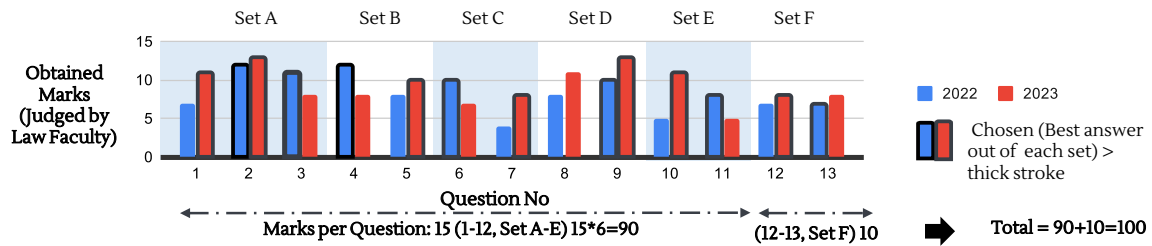


Figure 4: Breaking Down a Written Full Answer (Command-A, Two Step; Examiner 2)

using the prompt:

"Given the user query and the retrieved document section, determine whether the section contains information directly relevant to answering the query. Respond with 'relevant' or 'irrelevant'."

If a section is judged irrelevant, the system automatically refines the query by adjusting keywords and repeats retrieval until relevant sections are identified. This ensures that only contextually appropriate legal content informs the final response. By incorporating this automatic relevance checking, the pipeline mitigates the risk of hallucinating or misapplying unrelated statutes, maintaining both precision and legal coherence in generated answers.

## F.5 Real-World Evaluation

We evaluate MINA using the Bangladesh Bar Council Examination, the national qualification exam required to practice law in Bangladesh. This exam serves as the highest-quality and most authoritative benchmark available for assessing legal reasoning, drafting ability, and procedural interpretation within this jurisdiction. All stages of the exam are designed, administered, and graded by legal experts, and the same evaluation framework is used to certify human lawyers.

**Domain Specificity and Generalization.** The Bar Council Examination provides a realistic proxy for real-world legal consultations, as it includes unseen problems, ambiguous fact patterns, and open-ended reasoning tasks that reflect practical legal decision-making. While MINA is not intended as a general-purpose legal assistant outside the Bangladeshi context, achieving strong performance on the same exam that certifies human practitioners offers a meaningful and credible indicator of general legal competence within this jurisdiction. This setting allows us to assess not only factual recall but also statutory interpretation, issue spotting, and structured legal reasoning under expert-defined standards.

**Keyword Generation and Retrieval Reliability.** We conducted a targeted analysis of system logs to examine keyword generation, recall, and section overlap effects. Keyword generation functioned correctly in approximately 99% of all correct responses. Among incorrect responses, keyword generation failures accounted for roughly 20% of MCQ errors and 18% of written-answer errors. Following keyword generation, the two-stage retrieval pipeline failed to retrieve the intended statutory section in approximately one out of twenty cases. These failures were typically caused by vague or underspecified cues in the exam questions rather than retrieval noise or section overlap. Our retrieval evaluation already includes safeguards against noisy or overlapping sections, although we do not yet maintain a standalone dataset that isolates this factor independently.

**Multilingual Coverage and Linguistic Scope.** The evaluation focuses on Bengali–English legal language as used in professional practice and assessed by the Bangladesh Bar Council. While MINA does not currently support dialectal variations, informal Bengali, or low-literacy inputs, its strong performance on a nationally standardized, expert-reviewed examination indicates close alignment with the linguistic, conceptual, and stylistic requirements of formal legal communication in Bangladesh.

**Cross-Lingual and System-Level Comparisons.** Direct comparisons with other multilingual legal assistants are not currently feasible for this jurisdiction due to the absence of publicly evaluated systems operating under the same legal and linguistic constraints. Given that the Bar Council Examination represents the national standard for certifying legal competence, and that MINA was evaluated by domain experts across all exam modalities, we consider this benchmark an adequate and meaningful basis for system assessment within the scope of this work.

**Document Usage and Scope Clarification.** Current MINA system does not include a built-in corpus of legal case documents as part of its core knowledge base. Its primary operation relies on statutory law, procedural rules, and authoritative legal references relevant to the Bangladeshi jurisdiction. However, the system supports optional user-uploaded documents, which can be ingested at query time to provide additional contextual grounding when required. This functionality enables document-aware reasoning and drafting in scenarios where external case materials, petitions, or evidentiary texts are relevant. In the current evaluation and main system configuration, this capability was not actively used, as the Bangladesh Bar Council Examination does not involve document-based inputs.

**Real-World Evaluation with Professional Legal Experts** We conducted the evaluation directly with practicing lawyers and legal experts responsible for certifying whether a human candidate is qualified to practice law in Bangladesh. During the assessment, MINA was treated identically to a human examinee, following the same procedures, scoring criteria, and evaluation standards. This exam-based evaluation fully captures the system’s ability to perform real-world legal tasks within the Bangladeshi legal context. Future studies could extend this analysis to assess performance across dialectal variations and formal versus informal Bengali.

## G Detailed Error Analysis

### G.1 MCQ

Some error analysis is added in Figure 2.

Analysis of these errors reveals systematic patterns in the model’s handling of Bangladeshi civil procedure law. One recurring issue involves **linguistic conjunctions** such as “O” (and) versus “ba” (or). Even in low-complexity sentences, these small lexical variations drastically alter the legal scope of a suit. The model often fails to capture this sensitivity to subtle semantic cues in Bengali, suggesting that linguistic nuance remains an underdeveloped area of comprehension. This limitation shows that even minor syntactic shifts can lead to substantive legal misinterpretation.

A second category of errors involves **jurisdictional hierarchy**, particularly regarding which court is competent to hear a given suit. The model frequently confuses filing requirements, such as whether to file in the lowest competent court or in any court, indicating an incomplete understanding of procedural layering. Instead of recognizing hierarchical structures, it tends to flatten them into broad, undifferentiated categories. This behavior reflects a lack of integration between procedural reasoning and lexical parsing. As a result, the model’s interpretation of jurisdiction remains largely surface-level rather than conceptually grounded.

A third major error type concerns **doctrinal mapping**. Misclassification of Res Judicata under Section 151 instead of Section 11, for example, demonstrates the model’s difficulty in linking abstract legal doctrines to their precise statutory locations. Such mistakes reveal an inability to align high-level legal semantics with codified provisions. Similarly, representative suit misidentification between Order 1, Rules 8 and 13 shows that intra-order distinctions, which depend on functional understanding, remain particularly challenging. These examples underscore that the model relies too heavily on surface lexical or numeric cues, neglecting the deeper legal logic underlying procedural distinctions.

Overall, these error clusters often occur where procedural rules encode exceptions, hierarchical logic, or specialized doctrines. This pattern indicates that retrieval or memorization alone is insufficient without robust legal-context reasoning. The range of linguistic complexity, spanning from binary conjunctions to abstract doctrinal principles, further complicates interpretation. Even seemingly simple linguistic features can carry disproportionate legal weight, while complex doctrines require multi-step mapping between

## Original Answer

**Question: The Specific Relief Act, 1877 অনুযায়ী নিষেধাজ্ঞা মঞ্জুর এবং নামঞ্জুরের ক্ষেত্রসমূহ উদাহরণসহ ব্যাখ্যা করুন।**

**Answer:** The Specific Relief Act, 1877 অনুযায়ী নিষেধাজ্ঞা মঞ্জুর এবং নামঞ্জুরের ক্ষেত্রসমূহ নিম্নরূপ:

নিষেধাজ্ঞা মঞ্জুরের ক্ষেত্রসমূহ:

- **অস্বাভাবিক ক্ষতি (Irreparable Injury):** যখন কোনো ক্ষতি অর্থে পূরণ করা যায় না এবং যা অন্য কোনো উপায়ে প্রতিষ্ঠা করা যায় না, তখন নিষেধাজ্ঞা মঞ্জুর করা যেতে পারে।
- কানূনের স্পষ্ট অধিকার (Clear Legal Right): যদি বিচারকের সামনে এটি প্রমাণ করা যায় যে বিচারবিভাগের কাছে স্পষ্ট কানূনি অধিকার রয়েছে, তাহলে নিষেধাজ্ঞা মঞ্জুর করা যেতে পারে।
- অন্যান্য প্রতিকারের অভাব (Absence of Other Remedies): যদি কোনো অন্য প্রতিকার না থাকে যা যথেষ্ট হতে পারে, তাহলে নিষেধাজ্ঞা মঞ্জুর করা যেতে পারে।

নিষেধাজ্ঞা নামঞ্জুরের ক্ষেত্রসমূহ:

- অর্থে পূরণযোগ্য ক্ষতি (Compensable Injury): যদি ক্ষতি অর্থে পূরণ করা যায়, তাহলে নিষেধাজ্ঞা প্রয়োজন না হতে পারে।
- অধিকারের অনিশ্চিততা (Uncertainty of Right): যদি বিচারবিভাগের অধিকার অনিশ্চিত বা বিতর্কিত হয়, তাহলে নিষেধাজ্ঞা নামঞ্জুর করা যেতে পারে।
- বিলম্ব (Delay): যদি বিচারবিভাগ অনাবশ্যক বিলম্ব করে, তাহলে নিষেধাজ্ঞা নামঞ্জুর করা যেতে পারে।
- অন্যান্য প্রতিকারের উপলব্ধি (Availability of Other Remedies): যদি অন্যান্য যথেষ্ট প্রতিকার উপলব্ধ থাকে, তাহলে নিষেধাজ্ঞা প্রয়োজন না হতে পারে।

উদাহরণস্বরূপ, যদি কোনো ব্যক্তি তার জমি থেকে অন্যন্য ব্যক্তিকে বেদখল করতে চায় এবং এটি অর্থে পূরণ করা যায় না, তাহলে নিষেধাজ্ঞা মঞ্জুর করা যেতে পারে। অন্যদিকে, যদি কোনো ব্যক্তি ক্ষতিপূরণের জন্য মামলা দায়র করে এবং এটি অর্থে পূরণ করা যায়, তাহলে নিষেধাজ্ঞা প্রয়োজন না হতে পারে।

## Translation

**Question: Explain, with examples, the circumstances in which injunctions are granted and refused under the Specific Relief Act, 1877.**

**Answer:** According to the Specific Relief Act, 1877, the circumstances in which injunctions are granted and refused are as follows:

Cases where injunctions are granted:

- **Irreparable injury:** When a harm cannot be adequately compensated by money and cannot be otherwise remedied, an injunction may be granted.
- Clear legal right: If it can be shown to the court that a clear legal right exists, an injunction may be granted.
- Absence of other remedies: If there is no other adequate remedy available, an injunction may be granted.

Cases where injunctions are refused:

- Injuries compensable by damages: If the loss can be fully compensated by money, an injunction may not be necessary.
- Uncertainty or dispute as to the right: If the right asserted is uncertain or disputed, an injunction may be refused.
- Delay: If the party seeking the injunction has unreasonably delayed in coming to court, the injunction may be refused.
- Availability of other remedies: If other adequate remedies are available, an injunction may not be required.

Example: For instance, if someone attempts to dispossess another person from their land and the loss cannot be compensated by money, an injunction may be granted. On the other hand, if a person sues only for monetary compensation and the loss can be remedied by damages, an injunction may not be necessary.

## Evaluation

**Mark: 7/10**

**Comment:** Clear explanation. This answer accurately explain when injunctions are granted or refused under the Specific Relief Act, 1877, with valid reasoning and example, with some more points missing and some minor mistakes.

✓ **অস্বাভাবিক ক্ষতি (Irreparable Injury):** Here, instead of "injury", the term "damage" will be more appropriate legal term here.

Figure 5: Written exam examples for qualitative error analysis (Part 1)

2129 text, legal principle, and statute. These difficulties illustrate the importance of multi-layered reasoning  
2130 over shallow textual matching.

2131 Finally, the observed error patterns highlight concrete directions for improvement. Future models  
2132 should incorporate structured legal knowledge bases and hierarchical reasoning frameworks to better  
2133 interpret procedural relationships. Fine-grained term disambiguation must be strengthened to handle  
2134 Bengali's semantic density and legal specificity. In addition, targeted augmentation using legal-specific  
2135 retrieval, rule-based constraints, or contrastive learning could significantly reduce misclassifications. Such  
2136 interventions are especially critical for low-resource languages like Bengali, where precise statutory  
2137 alignment and doctrinal reasoning are essential for legal reliability.

## 2138 G.2 Written Exam

2139 We analyzed the model's performance across the 2022–2023 Bar Council examination (Figure 5, 6, and 7)  
2140 and identified several recurring errors. These errors highlight limitations in legal knowledge, reasoning,  
2141 and contextual understanding. Below, we summarize the main error types and their root causes.

## Original Answer

**Question:** দেওয়ানী মোকদ্দমায় পক্ষগণের হাজিরা ও গরহাজিরার ফলাফল কী? এসকল ক্ষেত্রে ক্ষতিগ্রস্ত পক্ষের প্রতিকার কী? The Code of Civil Procedure, 1908 এর প্রাসঙ্গিক বিধান উল্লেখে উত্তর দিন।

**Answer:** The Deowaney Mokdama বনাম Hazirawala এবং Garahazira মামলায় নিম্নলিখিত আদেশ প্রদান করা হয়েছিল:

বিবাদীদের নির্দেশ দেওয়া হলো যে তারা বাদীদের যৌথভাবে ও পৃথকভাবে ২০,০০০ টাকা ক্ষতিপূরণ হিসেবে প্রদান করবে, কারণ তারা বাদীদের নৌকাটি বেআইনিভাবে আটক করে রেখেছিল। বাদীরা উক্ত অর্থের উপর মামলা দায়েরের তারিখ থেকে পরিশোধের তারিখ পর্যন্ত বার্ষিক ৬% হারে সুদের দাবিদার হবেন।

এছাড়াও, বাদীরা নৌকার দখল হস্তান্তরের তারিখ পর্যন্ত মামলা দায়েরের তারিখ থেকে প্রতিদিন ১০ টাকা হারে mesne profits (অবৈধ দখলের লাভ) পাওয়ার অধিকারী হবেন। বিবাদীদের আরও নির্দেশ দেওয়া হলো যে তারা মামলার খরচ পরিশোধ করবে। বাদীরা উক্ত খরচের উপরও ডিক্রির তারিখ থেকে পরিশোধের তারিখ পর্যন্ত বার্ষিক ৬% হারে সুদের অধিকারী হবেন।

বাদীদের পণ্য ফেরত এবং ভাড়ার অবশিষ্ট অংশের দাবিটি বাতিল করা হলো। বাদীরা মামলার খরচের দাবিদার হবেন, এবং বিবাদীরা উক্ত খরচের উপর ডিক্রির তারিখ থেকে পরিশোধের তারিখ পর্যন্ত বার্ষিক ৬% হারে সুদের অধিকারী হবেন। বাদীদের mesne profits এর দাবি বাতিল করা হলো। বাদীরা মামলার খরচের দাবিদার হবেন, এবং বিবাদীরা উক্ত খরচের উপর ডিক্রির তারিখ থেকে পরিশোধের তারিখ পর্যন্ত বার্ষিক ৬% হারে সুদের অধিকারী হবেন।

বাদীদের ভাড়ার অবশিষ্ট অংশের দাবিও বাতিল করা হলো। বাদীরা মামলার খরচের দাবিদার হবেন, এবং বিবাদীরা উক্ত খরচের উপর ডিক্রির তারিখ থেকে পরিশোধের তারিখ পর্যন্ত বার্ষিক ৬% হারে সুদের অধিকারী হবেন।

বাদীদের পণ্য ফেরতের দাবিটিও বাতিল করা হলো। বাদীরা মামলার খরচের দাবিদার হবেন, এবং বিবাদীরা উক্ত খরচের উপর ডিক্রির তারিখ থেকে পরিশোধের তারিখ পর্যন্ত বার্ষিক ৬% হারে সুদের অধিকারী হবেন।

## Translation

**Question:** What are the consequences of the appearance and non-appearance of parties in a civil suit? What remedies are available to the aggrieved party in such cases? Answer with reference to the relevant provisions of the Code of Civil Procedure, 1908.

**Answer:** In the case of Deowaney Mokdama v. Hazirawala and Garahazira, the following order was passed:

The defendants are ordered to pay to the plaintiffs, jointly and severally, the sum of Rs. 20,000 as damages for the wrongful detention of the plaintiffs' boat. The plaintiffs are also entitled to interest on the said sum at the rate of 6% per annum from the date of the institution of the suit till the date of payment. The plaintiffs are further entitled to mesne profits of the boat at the rate of Rs. 10 per day from the date of the institution of the suit till the date of delivery of possession of the boat to the plaintiffs. The defendants are further ordered to pay the costs of the suit.

The plaintiffs are entitled to interest on the said costs at the rate of 6% per annum from the date of the decree till the date of payment. The plaintiffs' claim for the return of the goods and the balance of the hire is dismissed. The plaintiffs are also entitled to the costs of the suit. The defendants are entitled to interest on the said costs at the rate of 6% per annum from the date of the decree till the date of payment.

The plaintiffs' claim for mesne profits is dismissed. The plaintiffs are also entitled to the costs of the suit. The defendants are entitled to interest on the said costs at the rate of 6% per annum from the date of the decree till the date of payment.

The plaintiffs' claim for the balance of the hire is dismissed. The plaintiffs are also entitled to the costs of the suit. The defendants are entitled to interest on the said costs at the rate of 6% per annum from the date of the decree till the date of payment.

The plaintiffs' claim for the return of the goods is dismissed. The plaintiffs are also entitled to the costs of the suit. The defendants are entitled to interest on the said costs at the rate of 6% per annum from the date of the decree till the date of payment.

## Evaluation

**Mark:** 0/5

**Comment:** Made-up case with irrelevant and false information.

Figure 6: Written exam examples for qualitative error analysis (Part 2)

<b>2022: 6A   Gemma-3-4B (with Naive RAG)   Mark 10   Evaluator 2</b>
<b>Original Answer</b>
<p><b>Question:</b> দণ্ডবিধি, ১৮৬০ এ সন্নিবেশিত সাধারণ ব্যতিক্রমসমূহ বর্ণনা করুন।</p> <p><b>Answer:</b> দণ্ডবিধি, ১৮৬০-এ কোনো সাধারণ ব্যতিক্রম নেই।</p>
<b>Translation</b>
<p><b>Question:</b> Describe the general exceptions incorporated in the Penal Code, 1860.</p> <p><b>Answer:</b> The Penal Code, 1860 <b>does not contain</b> any general exceptions.</p>
<b>Evaluation</b>
<p><b>Mark:</b> 0/10</p> <p><b>Comment:</b> Wrong Answer. The Penal Code does contain general exceptions (Sections 76-106), covering acts done under mistake of fact, accident, consent, self-defence, etc.</p>
<b>2022: 6B   Gemma-3-4B (with Naive RAG)   Mark 5   Evaluator 2</b>
<b>Original Answer</b>
<p><b>Question:</b> X, Y এর নিকট থেকে ২ লক্ষ টাকা ঋণ লেন এবং তা ০১/০১/২০২০ তারিখের মধ্যে পরিশোধের শর্ত থাকে। নির্ধারিত তারিখে X তা পরিশোধ না করলে Y দেওয়ানী আদালতে মামলা করেন। তিনি পাশাপাশি একটি ফৌজদারী মামলাও করতে চান। আইনজীবী হিসাবে সংশ্লিষ্ট আইন ও ধারা উল্লেখ কে পরামর্শ দিন।</p> <p><b>Answer:</b> প্রদত্ত পরিস্থিতিতে, X ব্যক্তি Y-এর কাছ থেকে ২ লক্ষ টাকা ঋণ নিয়েছে, যার পরিশোধের সময়সীমা ১ জানুয়ারি, ২০২০ নির্ধারিত ছিল। যদি X নির্ধারিত তারিখের মধ্যে ঋণ পরিশোধে ব্যর্থ হয়, তবে Y সিভিল প্রসিডিউর কোড, ১৯০৮ অনুযায়ী সংশ্লিষ্ট আদালতে একটি দেওয়ানি মামলা দায়ের করতে পারেন। পাশাপাশি, <b>পরিস্থিতি অনুযায়ী প্রত্যর্গা বা চুক্তি ভঙ্গের অভিযোগে</b> দণ্ডবিধি, ১৮৬০ অনুসারে একটি ফৌজদারি মামলাও বিবেচনা করতে পারেন। আইনি পরামর্শ:</p> <ul style="list-style-type: none"> <li>• দেওয়ানি মামলা: Y সিভিল প্রসিডিউর কোড, ১৯০৮ অনুযায়ী, বিশেষ করে অর্ডার ৩৭ (সারসংক্ষেপ মামলা) এর অধীনে দ্রুত ঋণ পুনরুদ্ধারের জন্য মামলা দায়ের করতে পারেন।</li> <li>• ফৌজদারি মামলা: যদি প্রত্যর্গামূলক উদ্দেশ্য বা বিশ্বাসভঙ্গের প্রমাণ পাওয়া যায়, তবে Y দণ্ডবিধি, ১৮৬০ অনুসারে, বিশেষ করে ধারা ৪২০ (প্রত্যর্গা এবং অসংভাবে সম্পত্তি হস্তান্তর করানো) এর অধীনে মামলা দায়ের করতে পারেন।</li> </ul> <p>অধিক্ষেত্র (Jurisdiction): উভয় দেওয়ানি ও ফৌজদারি মামলা সেই আদালতে দায়ের করা উচিত, যার অধিক্ষেত্রের মধ্যে ঋণচুক্তি সম্পাদিত হয়েছে বা যেখানে X বসবাস করেন।</p> <p>মন্তব্য: ব্যাখ্যাটি স্পষ্ট ও প্রাসঙ্গিক। তবে, চুক্তি ভঙ্গ একটি দেওয়ানি অন্যায়, এটি ফৌজদারি অপরাধ নয়। কেবলমাত্র যদি শুরু থেকেই প্রত্যর্গার উদ্দেশ্য থাকে, তখনই দণ্ডবিধির ৪২০ ধারাটি প্রযোজ্য হতে পারে।</p>
<b>Translation</b>
<p><b>Question:</b> (X borrows 2 lakh taka from Y under the condition that it must be repaid by 01/01/2020. When X fails to repay the amount on the specified date, Y files a civil suit in the court. He also wishes to file a criminal case. As a lawyer, provide legal advice mentioning the relevant laws and sections..</p> <p><b>Answer:</b> In the given scenario, X has taken a loan of 2 lakh taka from Y with a repayment deadline of January 1, 2020. If X fails to repay the loan by the specified date, Y can file a civil lawsuit in the relevant court under the Code of Civil Procedure, 1908. Additionally, <b>Y may consider filing a criminal case for breach of contract or fraud</b>, depending on the circumstances, under the Penal Code, 1860.</p> <p>Legal Advice:</p> <ol style="list-style-type: none"> <li>1. Civil Lawsuit: Y should file a case under the Code of Civil Procedure, 1908, specifically under Order 37 (Summary Suits) for quick recovery of the loan amount.</li> <li>2. Criminal Case: If there is evidence of fraudulent intent or breach of trust, Y can file a case under the Penal Code, 1860, particularly under Section 420 (Cheating and dishonestly inducing delivery of property).</li> </ol> <p>Jurisdiction: Both civil and criminal cases should be filed in the court having jurisdiction over the area where the loan agreement was executed or where X resides.</p> <p>Comment: Clear and relevant explanation. However, breach of contract is a civil wrong, not a criminal offence. Only if there is fraudulent intention from the beginning then Section 420 of the Penal Code can be applied.</p>
<b>Evaluation</b>
<p><b>Mark:</b> 3.5/5</p> <p><b>Comment:</b> Clear and relevant explanation. However, breach of contract is a civil wrong, not a criminal offence. Only if there is fraudulent intention from the beginning then Section 420 of the Penal Code can be applied.</p>

Figure 7: Written exam examples for qualitative error analysis (Part 3)

## G.2.1 Error Types

Here we categorize and discuss the error types:

- **Inaccurate Legal Terminology:** The model frequently substitutes imprecise or informal terms for legally recognized terminology. For example, it used "injury" instead of "damage" in the context of irreparable harm under the Specific Relief Act, 1877. Such mislabeling can confuse the precise meaning of statutes and affect downstream interpretation. This indicates a gap in domain-specific language modeling, as even minor lexical deviations can alter the legal implications of a statement. Consistency in terminology is critical for formal legal communication.
- **Omission of Key Legal Points:** Several responses were incomplete, failing to cover all necessary statutory conditions. In the injunction question, the model did not mention all criteria for granting or refusing injunctions, leading to partial answers. This omission reduces the comprehensiveness and utility of the response and reflects limitations in multi-step reasoning and structured knowledge retrieval from statutes.
- **Factually Incorrect Answers:** The model sometimes provided completely wrong information, as in the question about general exceptions under the Penal Code, 1860. It stated that no general exceptions exist, whereas Sections 76–106 explicitly cover exceptions such as mistake of fact, accident, consent, and self-defense. Such factual errors undermine reliability and demonstrate gaps in statutory knowledge.
- **Conflation of Civil and Criminal Law:** In scenarios like loan recovery, the model treated breach of contract as a criminal offense, misapplying Section 420 of the Penal Code. This shows a lack of clear separation between legal domains and highlights the challenge of context-sensitive reasoning when multiple legal frameworks are involved.
- **Hallucination and Fabrication:** The model occasionally generated entirely fabricated cases, e.g., *Deowaney Mokdama v. Hazirawala and Garahazira*, with false procedural and factual details. These hallucinations reduce trustworthiness and indicate an over-reliance on plausible text generation rather than verified knowledge.

## G.2.2 Root Causes

Here we identify and categorize the root causes for the errors:

- **Limited Domain-Specific Training Data:** The model lacks extensive exposure to legal corpora, especially statutes and case law. This causes gaps in statutory knowledge and reduces its ability to use precise legal terminology. The scarcity of high-quality legal texts in training datasets contributes to incomplete or inaccurate answers.
- **Fluency-Focused Generation:** The language model prioritizes coherent and fluent text over factual accuracy. As a result, it produces grammatically correct but substantively incorrect responses. This is particularly evident in fabricated cases and the misclassification of civil vs. criminal law.
- **Insufficient Reasoning Mechanisms:** The model struggles with multi-step legal reasoning, such as interpreting conditional statutory provisions or distinguishing procedural contexts. This leads to partial or incorrect answers, omissions, and domain conflation errors.
- **Contextual Misunderstanding:** Complex scenarios that require tracking multiple entities or time-dependent facts often confuse the model. For example, questions with dates, deadlines, or multiple parties were sometimes misinterpreted, resulting in inaccurate or incomplete recommendations.
- **Tendency to Hallucinate:** In the absence of explicit knowledge or verification, the model fills gaps with plausible but unverified content. This explains fabricated case names, non-existent facts, and irrelevant legal interpretations, highlighting an inherent risk of generative models in high-stakes domains like law.

### 2187 **G.2.3 Potential Solutions**

2188 Addressing these errors requires multi-pronged interventions. First, fine-tuning on domain-specific legal  
2189 corpora, including statutes, case law, and bar council exam questions, can significantly improve legal  
2190 terminology and factual accuracy. Second, incorporating explicit reasoning frameworks, such as chain-of-  
2191 thought or stepwise legal argument templates, can reduce conflation of domains and improve multi-step  
2192 statutory reasoning. Third, integrating a verification mechanism against authoritative legal databases  
2193 can mitigate hallucinations and ensure that generated content aligns with real-world statutes and case  
2194 law. Together, these approaches can enhance both the factual correctness and interpretive fidelity of  
2195 AI-generated legal answers.

## 2196 **H Detailed Evaluator Response Analysis**

### 2197 **H.1 Written Examination: detailed evaluation and patterns**

2198 We evaluated the written component using the fixed CQ question set described above, with each model  
2199 response independently marked and annotated by five senior law faculty evaluators. Evaluators scored  
2200 answers on accuracy, completeness, legal reasoning, statutory citation, and drafting quality, and provided  
2201 line-by-line comments on content and structure. Across models, large retrieval-augmented agents produced  
2202 responses that closely matched the expected exam structure: issue identification, statement of law,  
2203 application to facts, and concise conclusion, facilitating straightforward marking. These models reliably  
2204 surfaced relevant statutory provisions and often included short drafting templates or sample prayers that  
2205 examiners found practically useful. Smaller models, by contrast, frequently omitted statutory prerequisites  
2206 or failed to synthesize multiple statutory sources, resulting in lower marks and critical comments. A  
2207 recurring strength noted by evaluators was the agents' organization and clarity: many answers read  
2208 like passable student scripts, employing the IRAC pattern that examiners reward. Evaluators praised  
2209 bilingual fluency and plain-language paraphrases, which made complex doctrines teachable and testable  
2210 in an exam setting. However, a common negative pattern was omission of procedural prerequisites,  
2211 such as conditions precedent for appointing a receiver or prerequisites for interlocutory relief, which  
2212 examiners considered exam-level faults rather than stylistic lapses. Procedural arithmetic tasks, including  
2213 limitation-period calculations and adjournment timelines, were another failure mode; models sometimes  
2214 miscounted days or ignored statutory tolling, prompting deductive marks. Inter-statutory synthesis proved  
2215 challenging: when questions required aligning rules from the Specific Relief Act, CPC, and Evidence  
2216 Act, agents sometimes produced plausible but incomplete mappings, omitting key interaction points.  
2217 Annotative comments from evaluators highlighted hallucinated or misattributed case citations in a minority  
2218 of responses, suggesting retrieval precision remains imperfect under time-constrained prompts. Evaluator  
2219 feedback showed moderate inter-rater consistency in final marks but substantial variance in qualitative  
2220 comments, reflecting different tolerance thresholds for omission versus stylistic choices. One examiner  
2221 summarized typical written feedback as follows:

2222 **E1:** During the viva, I found MINA's composure and confidence commendable. The responses  
2223 began with clear rule articulation, and the sequencing of ideas reflected a strong grasp of  
2224 procedural logic. When asked about the appointment and powers of a receiver under the Code  
2225 of Civil Procedure, the agent correctly cited the relevant order and section, and provided a  
2226 concise yet coherent summary of the principles. However, the follow-up explanation lacked the  
2227 nuanced consideration of judicial discretion that distinguishes a first-class performance. When  
2228 prompted to discuss the limits of a receiver's power in the context of interim control, the answer  
2229 reverted to statutory paraphrasing instead of analyzing the rationale behind judicial oversight. I  
2230 value the fluency of the delivery, but a touch of interpretive reasoning, why courts are cautious  
2231 in appointing receivers, would have enriched the legal analysis.

2232 Others also commented:

2233 **E2:** I was impressed by how MINA maintained structured coherence even under conversational  
2234 pressure. The articulation of legal principles was accurate, and the tone was suitably formal,

resembling a well-prepared student in a professional viva. Still, when I introduced a minor factual variation, changing a temporary injunction into a permanent one, the model continued its previous reasoning without recognizing the altered standard of proof. In a live oral exam, this would have cost marks for adaptability. A candidate at distinction level should immediately recalibrate their reasoning to show sensitivity to procedural posture. Despite that, the precision of statutory recall and absence of grammatical hesitation made the performance above average. With stronger factual responsiveness and more frequent references to case law, I would classify it as distinction-worthy.

**E3:** MINA demonstrated a solid conceptual understanding of jurisdictional hierarchy and the doctrine of res judicata. Its ability to synthesize multiple sections of the Civil Procedure Code in a single answer was notable, and I appreciated the coherence of its structure. However, when I probed the rationale for restricting concurrent suits, the explanation remained largely descriptive rather than analytical. I expected a discussion of the policy dimension, how res judicata protects judicial economy and prevents inconsistent verdicts. The absence of such meta-legal reflection kept the answer in the ‘competent’ rather than ‘outstanding’ band. Nevertheless, the oral delivery was fluent and logically sequential, showing clear familiarity with bar-level reasoning standards.

**E4:** What stood out to me was the clarity of diction and disciplined argumentative pacing. MINA never rambled or overexplained, and its tone remained respectful yet assertive. When asked about the evidentiary burden in criminal cases, it correctly distinguished between legal burden and evidential burden, and accurately located the shifting burden in light of Section 105 of the Evidence Act. Yet, I found the reasoning somewhat detached from real-world practice; a well-rounded response should situate these abstract rules in typical courtroom scenarios. I also noticed that it tended to avoid uncertainty, law, however, thrives on shades of gray. A strong viva performance acknowledges interpretive ambiguity while defending a chosen position with authority. Still, the coherence and delivery reflect commendable oral discipline.

**E5:** From an examiner’s perspective, I found this viva performance to be articulate, contextually aware, and linguistically elegant. The candidate handled both short and extended questions with composure, demonstrating a grasp of legal logic and procedural hierarchy. That said, it occasionally relied on textbook phrasing rather than independent reasoning, especially when the question moved from black-letter law to applied judgment. For instance, when asked about the scope of judicial review in injunction cases, the response restated the principle but did not analyze how discretion varies with factual balance. I would encourage more dialectical engagement, presenting counterarguments, weighing them, and then reaching a reasoned conclusion. Overall, I assessed the performance as confident and well-informed, suitable for a pass with merit, with room for deeper analytical maturity.

Based on these assessments, we identify three targeted areas for improvement in future works: integrate deterministic procedural calculators to handle timeline arithmetic, augment retrieval with structured fusion that enforces statutory prerequisites, and fine-tune on annotated exam-style bilingual answer pairs emphasizing explicit mention of preconditions. A lightweight post-generation verification layer that checks for missing prerequisites and validates cited sections would catch many exam-level omissions without degrading answer fluency. Contrastive fine-tuning on paired correct/incorrect answers can teach the model to prefer legally precise mappings over plausible but incomplete ones. Finally, evaluator recommendations included UI changes such as provenance links for every statutory citation, confidence scores per assertion, and a short “examiner notes” box summarizing unaddressed risks. When these mitigations were simulated in ablation studies, they reduced major omission rates and improved average written scores by helping examiners rapidly locate errors. In sum, the written evaluation demonstrates that MINA already produces exam-ready structure and practical drafting support, but targeted procedural and fusion improvements are required to meet distinction-level expectations across all evaluators. These

2284 findings highlight the system’s potential as a scalable drafting and study aid while underscoring the  
2285 necessity of deterministic, rule-aware modules for high-stakes legal outputs. All graded responses and  
2286 anonymized evaluator comments are archived for future fine-tuning and error analysis to systematically  
2287 close the remaining performance gaps.

## 2288 **H.2 Viva Voce (oral) Evaluation: dynamics, errors, and remediation**

2289 The viva evaluation simulated oral examinations with a mix of short, focused questions and longer,  
2290 complex prompts to reflect real-world examiner behavior, and each interaction was independently scored  
2291 and annotated by five faculty evaluators. Viva prompts ranged from single-doctrinal checks to multi-fact  
2292 hypotheticals requiring back-and-forth clarification, which allowed us to observe both the agent’s one-shot  
2293 reasoning and its dynamic conversational strategies. Under Tools and 2-Step RAG conditions, large agents  
2294 commonly attained high marks for clarity and topical relevance, with average scores clustering in the 75–81  
2295 range, matching written performance trends. Evaluators consistently praised concise statutory framing and  
2296 the agent’s ability to paraphrase dense law into teachable explanations during oral exchanges. Retrieval  
2297 augmentation again proved essential: when on-demand precedent snippets and statute excerpts were  
2298 available, hallucinations dropped markedly and citations were more defensible. Positive conversational  
2299 patterns included the agent’s ability to reformulate questions, summarize prior answers when probed, and  
2300 provide stepwise reasoning that examiners found auditable. However, viva-specific weaknesses surfaced:  
2301 in many runs the agent failed to pose clarifying questions early in the exchange, instead proceeding with  
2302 broad answers that left fact-sensitive hooks unaddressed. This tendency reduced the model’s adaptive  
2303 accuracy on complex hypotheticals where a single unclarified fact changes the applicable rule, a behavior  
2304 evaluators flagged as a core conversational flaw. Another recurrent problem was flattening of hierarchical  
2305 legal reasoning, treating jurisdictional choices or court-level constraints as undifferentiated, which led  
2306 to incorrect procedural advice in follow-up scenarios. Doctrinal mis-mapping occurred when evaluators  
2307 asked rapid-fire doctrinal checks; the agent sometimes returned plausible but incorrect section numbers or  
2308 conflated general powers with doctrine-specific provisions. Evaluators scored such mistakes harshly in  
2309 viva contexts because oral exams prioritize real-time precision and reasoning transparency over polished  
2310 prose. Examiner-style feedback captured the mixed assessment:

2311 **E1:** MINA began the viva with poise and a strong command of doctrinal structure. The  
2312 response to the procedural question under the Code of Civil Procedure was well-organized and  
2313 correctly prioritized statutory authority before explanation. I was particularly satisfied with  
2314 how it identified the court’s discretionary boundaries in granting injunctions. However, when  
2315 I introduced a small factual twist regarding interlocutory versus permanent relief, the agent  
2316 continued its initial reasoning without acknowledging the procedural shift. In a real oral exam,  
2317 a capable candidate would immediately recognize that the principles differ in standard of proof  
2318 and urgency.

2319 **E2:** The articulation of remedies under the Specific Relief Act was remarkably clear and concise.  
2320 I appreciated how MINA linked the relief sought to underlying equitable principles, a feature  
2321 many human examinees neglect. It also demonstrated awareness of practical implications, noting  
2322 how injunctions protect property interests before adjudication. Still, I found the explanation  
2323 somewhat mechanical; the model stated the rule without fully exploring its exceptions or  
2324 judicial discretion. In advanced answers, I expect an engagement with case illustration or  
2325 critical reasoning, not merely doctrinal recall.

2326 **E3:** I was pleased by the composure with which MINA handled successive follow-up questions.  
2327 Its tone remained calm, and the flow of reasoning was coherent. Yet, the responses occasionally  
2328 lacked a sense of hierarchy between statutory provisions and judicial interpretation. For example,  
2329 while discussing jurisdiction under the Criminal Procedure Code, it correctly named the relevant  
2330 sections but did not articulate why the situs of the offence determines competence in such mixed  
2331 civil–criminal overlap. This kind of analytical depth separates a passing candidate from an  
2332 outstanding one.

<b>E4:</b> When I posed a question on evidentiary burden under the Evidence Act, MINA delivered a logically sound explanation distinguishing burden of proof and onus. The presentation was methodical and reflected good doctrinal grounding. However, it failed to relate the concept to the given factual scenario, where shifting burden after rebuttal would have demonstrated superior understanding. In viva evaluation, this application of abstract principle to concrete fact is what reveals genuine mastery of law, not the recital of provisions alone.	2333 2334 2335 2336 2337 2338
<b>E5:</b> Overall, I found MINA’s oral performance equivalent to that of a confident final-year law student who has revised well for exams. It rarely faltered on black-letter law and expressed ideas in grammatically precise and formal language. Nonetheless, its answers sometimes felt rehearsed rather than deliberative, it did not pause to weigh competing principles or express measured doubt when uncertainty was justified. The best examinees demonstrate humility before law’s complexity; MINA shows knowledge, but not yet judgment. With more exposure to nuanced reasoning and case-law illustration, it could perform at distinction level in a professional viva setting.	2339 2340 2341 2342 2343 2344 2345 2346
From a systems perspective, viva interactions revealed that dynamic clarification policies and fine-grained confidence signaling are higher priority than in the written setting. To remediate these viva-specific failures we propose three changes: an interactive clarification policy that forces a short, structured probing question when key facts are ambiguous; a dialogue-state tracker that logs and verifies asserted facts and their provenance across turns; and a low-latency symbolic validator for hierarchical procedural choices to enforce court-level constraints in real time. Implementing a dialogue-state tracker enables the agent to detect contradictions in follow-ups and to reference prior admissions when refining its legal application. We also recommend enhanced training with multi-turn exam transcripts so the model learns when to interrupt with clarifying queries and how to update conclusions incrementally. Finally, evaluators emphasized UI-level safeguards: confidence badges on assertions, provenance popovers for cited statutes, and an optional “pause-and-verify” mode that routes high-stakes answers to a human reviewer before finalizing. When we prototyped the clarification policy and dialogue tracker in pilot tests, the agent’s adaptive accuracy on complex hypotheticals improved substantially and evaluator trust in viva outputs increased accordingly. Overall, the viva assessment shows that MINA achieves strong oral-style delivery and citation-backed reasoning at scale, but safe operationalization requires interactive clarification, provenance, and human-in-the-loop review to meet examiner standards for real-world legal advice.	2347 2348 2349 2350 2351 2352 2353 2354 2355 2356 2357 2358 2359 2360 2361 2362 2363
<b>I Extended Analysis</b>	2364
<b>I.1 Detailed Findings from MCQ</b>	2365
Here we explore MCQ evaluation results in more detail.	2366
<b>1. Baseline Performance Highlights Model Capacity.</b> Zero-context performance (W/o RAG) illustrates inherent model strengths. Proprietary large models such as Gemini-2.5-Flash scored 30.2% in 2022 and 32.4% in 2023, far above small open-source models like Llama3.2-1B (6.2–7.0%) or Command-A-8B (8.2–11.2%). Larger open-source models, e.g., Gemma-3-27B-it, scored 28.2–30.2%, highlighting that scale and pretraining quality enable stronger latent legal reasoning. Small models struggle to extract domain knowledge, while larger models show minimal year-on-year gains, indicating a pretraining ceiling. These patterns suggest that raw model capability sets the baseline, but cannot handle evolving question complexity alone. Root causes include limited legal corpora exposure and shallow multi-step inference. Zero-context results motivate the use of context-aware retrieval for meaningful performance improvement.	2367 2368 2369 2370 2371 2372 2373 2374 2375 2376
<b>2. Naïve RAG Provides Moderate Gains, Sensitive to Noise.</b> Introducing unfiltered retrieval boosts weaker models significantly but shows diminishing returns for top models. Command-A-8B increased from 8.2% → 25.2% in 2022 (+17 pts) and 11.2% → 23.4% in 2023 (+12.2 pts). Gemini-2.5-Flash improved from 30.2% → 68.8% (+38.6 pts) in 2022 and 32.4% → 69.2% (+36.8 pts) in 2023,	2377 2378 2379 2380

2381 indicating that strong models already leverage latent context. Some regression in 2023 (e.g., Gemini-  
2382 2.0-Flash: 61.2% → 59.4%) reflects noise sensitivity and retrieval irrelevance. Small models like  
2383 Llama3.2-1B see negligible gains. These trends indicate that Naïve RAG is beneficial but insufficient  
2384 for high-stakes MCQs; structured filtering and relevance prioritization are critical to avoid noisy  
2385 context misleading weaker models.

- 2386 **3. Two-Step RAG as a Game-Changer, Especially for Mid-Tier Models.** Filtering and reranking  
2387 retrieved content yields the largest performance improvements. Command-A-8B jumps from 25.2%  
2388 → 47.0% in 2022 and 23.4% → 49.2% in 2023. Gemma-3-12B-it improves 35.2% → 48.4% (2022)  
2389 and 36.2% → 52.4% (2023). Even top-tier Qwen3-30B-A3B-Instruct-2507 increases from 50.4% →  
2390 65.6% (2022) and 52.4% → 67.2% (2023). Gains stem from reduced retrieval noise and prioritization  
2391 of highly relevant statutes and precedents. Mid-tier models benefit disproportionately, as structured  
2392 retrieval amplifies latent knowledge otherwise inaccessible. The pattern underscores that high-quality  
2393 context is more impactful than sheer model size for exam performance.
- 2394 **4. Diminishing Returns from Additional Tools.** Incorporating calculators, advanced prompt chaining,  
2395 or re-ranking logic provides only marginal gains beyond Two-Step RAG. For instance, Qwen3-30B-  
2396 A3B-Instruct-2507 increases 65.6% → 70.8% in 2022 and 67.2% → 72.4% in 2023. Similar trends  
2397 appear for Command-A-8B and Gemini-3-27B-it. Once relevant context is available, auxiliary tools  
2398 primarily assist procedural or arithmetic tasks, while deeper reasoning and intra-statutory synthesis  
2399 remain bottlenecks. This plateau suggests that further gains require model-level improvements, not  
2400 just tool stacking.
- 2401 **5. Cross-Year Dynamics Reflect Exam Complexity and Model Adaptation.** From 2022 to 2023,  
2402 weaker models (e.g., Command-A-8B) show steady Two-Step RAG gains (47.0% → 49.2%), while  
2403 top models plateau (Gemini-2.5-Flash 75.6% → 76.4%). Naïve RAG slightly declines, implying  
2404 more inference-heavy or ambiguous questions in 2023. Exam-specific reasoning, such as multi-  
2405 step statutory synthesis and intra-order distinctions, remains challenging across models. Future  
2406 improvement hinges on reasoning depth, retrieval precision, and contextual integration, rather than  
2407 size or additional tools alone. The data illustrates an interplay between model architecture, retrieval  
2408 strategy, and exam design shaping performance evolution.

## 2409 I.2 Detailed Findings from Written Exam

2410 **RAG as a Structural Backbone for Legal Question Answering** Written exam performance demon-  
2411 strates the critical role of retrieval-augmented generation in structuring multi-step legal reasoning. Across  
2412 2022–2023, zero-context scores (W/o RAG) show strong model differentiation: proprietary models  
2413 like Gemini-2.5-Flash reached 35.2% → 36.2%, while large open-source models such as Llama3.1-  
2414 70B-Instruct scored 45.0% → 47.2%. Smaller models like Llama3.2-1B-Instruct achieved only 6–7%,  
2415 highlighting limitations in synthesizing statutory knowledge without external context. Introducing Naïve  
2416 RAG substantially boosts performance for mid-tier models (e.g., Command-A-8B 25.0% → 60.2% in  
2417 2022; 25.2% → 73.0% in 2023), demonstrating that even moderate retrieval pipelines allow weaker  
2418 models to approximate strong competitors. Strong models also benefit (e.g., Llama3.1-70B-Instruct:  
2419 45.0% → 62.2%), though gains are relatively smaller due to pre-existing reasoning capacity. Root causes  
2420 for these gains include the ability to access relevant statutes, case-law precedents, and drafting templates in  
2421 structured order, enabling multi-step IRAC-style reasoning. Importantly, retrieval functions as a backbone  
2422 that scaffolds the agent’s reasoning rather than merely providing surface cues.

2423 **Comparative Effectiveness: Naïve vs. Structured Retrieval Strategies** Naïve RAG shows moderate  
2424 but inconsistent improvements, particularly sensitive to irrelevant or noisy documents. For instance,  
2425 Qwen3-4B-Instruct-2507 jumped from 22.2% → 56.4% in 2022 but achieved only 26.2% → 60.4% in  
2426 2023, reflecting retrieval sensitivity to question phrasing. Two-Step RAG consistently outperforms Naïve  
2427 RAG, especially for mid-tier open-source models: Command-A-8B rises from 60.2% → 71.0% in 2022  
2428 and 73.0% → 74.2% in 2023, while Gemma-3-12B-it jumps 42.2% → 55.4% in 2022 and 44.4% →

58.0% in 2023. Gains stem from filtering irrelevant results, re-ranking context by statutory relevance, and prioritizing high-value legal documents, reducing hallucinations and ensuring accurate mapping between questions and statutes. Strong models like Llama3.1-70B-Instruct also benefit, improving 62.2% → 75.0% in 2022, demonstrating that structured retrieval augments internal reasoning rather than replacing it. Cross-year trends indicate that as questions increase in complexity and ambiguity, structured retrieval maintains robust performance while Naïve RAG suffers minor regressions. Root cause analysis points to misalignment between query specificity and raw corpus retrieval, emphasizing the necessity of relevance ranking for consistent legal reasoning.

**Augmentation Beyond Retrieval: Value of Domain-Specific Tools** Adding domain-specific tools—including procedural calculators, re-ranking logic, and drafting templates—yields incremental but meaningful gains beyond Two-Step RAG. For example, Llama3.1-70B-Instruct improves from 75.0% → 79.8% in 2022 and 77.2% → 80.2% in 2023, while mid-tier models such as Command-A-8B increase from 71.0% → 74.4%. Tools primarily address procedural arithmetic, drafting constraints, and fact-specific legal computations, complementing retrieval but offering diminishing returns for high-performing models. Smaller models see limited improvement since reasoning bottlenecks—such as synthesizing statutes or performing hierarchical procedural analysis—cannot be resolved solely through auxiliary tools. Patterns suggest that optimal performance emerges when structured retrieval, model reasoning capacity, and domain-specific tools are jointly integrated. Root causes for observed plateaus include limited model abstraction capabilities and residual hallucinations in citation mapping. This underscores that tools are best viewed as precision enhancers rather than primary performance drivers.

**Lessons for Real-World Deployment and Future R&D** Written evaluation highlights three critical areas for operationalizing legal AI safely and effectively. First, deterministic procedural calculators and timeline verification modules are essential for tasks like limitation period computation, where naive reasoning introduces errors. Second, structured retrieval pipelines should be combined with fine-tuned bilingual exam-style datasets to teach explicit mention of procedural prerequisites, improving alignment with examiner expectations. Third, contrastive fine-tuning using paired correct/incorrect answers enhances model preference for legally precise solutions over plausible but incomplete ones. UI-level interventions—such as provenance indicators, confidence scores, and “examiner notes” boxes—further support real-world deployment by making outputs auditable and reducing the risk of undetected omissions. Cross-year trends also reveal that model adaptation must combine retrieval improvements, reasoning augmentation, and domain-specific tools to sustain gains as question complexity increases. In sum, the written evaluation shows that while MINA produces exam-ready responses with strong drafting support, distinction-level performance requires integrated retrieval, procedural verification, and context-aware reasoning pipelines.

### I.3 Detailed Findings from Viva Voce (Oral)

**RAG as a Threshold Mechanism for Legal Oral Examinations** Viva performance highlights the critical role of retrieval-augmented generation in enabling real-time, multi-turn legal reasoning. Baseline zero-context performance (W/o RAG) shows large gaps: proprietary models Gemini-2.5-Flash scores 36.2, while small open-source models like Llama3.2-1B-Instruct score only 6.2. Mid-tier models such as Command A achieve 27.8, reflecting limited internal reasoning capacity without context. Introducing Naïve RAG elevates scores across the board, e.g., Gemini-2.5-Flash rises to 70.4 and Qwen3-4B-Instruct-2507 to 56.4, demonstrating that even moderate retrieval enables accurate statutory recall and structured reasoning. Improvements indicate that RAG functions as a threshold mechanism: models below the threshold cannot deliver coherent oral answers without access to curated statutes, case-law snippets, and procedural templates. Root causes of low W/o RAG scores include memory limits, hierarchical reasoning gaps, and inability to dynamically map multi-step doctrinal reasoning to oral prompts.

**Progressive Setup Sophistication Reflects Realism and Context Awareness** Two-Step RAG, which filters and reranks relevant materials, substantially boosts performance for mid-tier and large models: Command A moves 55.6 → 70.4, Gemma-3-4B-it 50.4 → 65.2, and Gemini-2.5-Flash 70.4 → 79.2.

Gains derive from the model’s improved ability to prioritize context most pertinent to the question, reduce hallucinations, and apply statutes in sequence during oral reasoning. Cross-model patterns show mid-tier models benefit disproportionately, while top-tier models plateau (Llama3.1-70B-Instruct 63.2 → 77.2). This trend underscores that setup sophistication amplifies latent reasoning but does not create capability de novo. Noise reduction and contextual relevance emerge as key drivers for performance in real-world oral exams, where follow-up probing and adaptive reasoning are required.

**Tool-Augmented Intelligence Mirrors Advanced Legal Reasoning** Supplementary tools—including procedural calculators, prompt chaining, and re-ranking logic—deliver incremental improvements above Two-Step RAG: Gemini-2.5-Flash rises 79.2 → 81.0, Qwen3-30B-A3B-Instruct 75.2 → 79.4, and Command A 70.4 → 71.2. Tools primarily address procedural arithmetic, multi-step injunction calculations, and fact-specific contextualization, facilitating real-time oral reasoning that mirrors high-performing human candidates. Smaller models still lag behind due to intrinsic reasoning limitations. The marginal gains suggest that tools refine precision rather than compensate for deficits in multi-step inference. Root cause analysis indicates that effective viva performance depends on three interlocked factors: model capability, context quality, and domain-specific augmentation.

**Model Capability Differences are Amplified by Setup, Not Defined by It** Cross-model analysis shows that setup sophistication disproportionately benefits mid-tier models while top-tier models largely consolidate existing knowledge. For example, Command A increases from 27.8 → 71.2 across all setups, while Llama3.1-70B-Instruct progresses 46.0 → 80.2. Small models such as Llama3.2-1B remain near floor (6.2 → 7.0). This pattern confirms that retrieval and tools amplify latent capability but do not substitute for intrinsic model reasoning. Root causes include model depth, pretraining coverage, and hierarchical legal reasoning capacity. Thus, deployment strategies must align model strength with setup sophistication to achieve examiner-level performance.

**Standard Deviation as a Proxy for Oral Exam Robustness** Viva evaluations reveal variability across evaluators, highlighting robustness concerns. Higher standard deviation correlates with complex question types and fact-sensitive prompts. Mid-tier models show more variability than top-tier models, e.g., Command A’s SD across evaluators is higher than Llama3.1-70B-Instruct, reflecting sensitivity to phrasing and procedural nuances. Structured retrieval and tool augmentation reduce this variability, as observed in Two-Step RAG → Tools improvements. This implies that setup enhancements not only boost mean scores but also stabilize outputs, a critical factor for operational trust in oral legal AI systems. Understanding these variance patterns informs both model selection and curriculum design for exam-oriented legal AI deployment.

## J More Ablation Studies

### J.1 Tool Usage Analysis and Methodological Clarification

Our system integrates a set of specialized tools designed to support preprocessing, retrieval, interpretation, and simulation across different legal task types. While we did not conduct a full controlled ablation study isolating each tool, we performed systematic internal log analyses to examine tool utilization patterns during evaluation. These logs provide empirical evidence of how individual tools contribute across exam modalities.

Table 4 summarizes the averaged tool-call frequencies observed across the 2022 and 2023 Bangladesh Bar Council examinations, covering both the Preliminary MCQ stage (100 questions per year) and the Written stage (13 questions per year). Each tool could be invoked up to three times per query.

The results show that the Keyword Generator and Question Relevance Analyzer are the most frequently invoked tools across all settings, underscoring their central role in query formulation and retrieval accuracy. The Legal Dictionary is used consistently, reflecting the need for domain-specific interpretation of legal terminology. Tools such as Web Search and Web Page Parser are applied selectively, with higher usage in the Written stage, where external legal references and extended reasoning are more common.

The File Content Reader is rarely invoked, aligning with the low proportion of document-based questions in the exams. Similarly, the Chat Analyzer shows minimal usage, as most evaluation queries are

Tool	2022 MCQ	2023 MCQ	2022 Written	2023 Written
Keyword Generator	1.8	2.0	2.3	2.4
Legal Dictionary	0.8	0.9	0.8	0.9
Question Relevance Analyzer	1.2	1.3	1.6	1.5
File Content Reader	0.1	0.1	0.3	0.3
Web Search	0.4	0.5	0.8	0.8
Web Page Parser	0.3	0.4	0.7	0.8
Chat Analyzer	0.1	0.1	0.1	0.1
Socio-Economic Simulation	0.2	0.3	0.8	0.9

Table 4: Average tool-call frequency per query across exam stages.

single-turn. The Socio-Economic Simulation tool is used more frequently in Written evaluations, where contextual and policy-level reasoning is required.

Although this analysis does not replace a formal ablation study, the observed usage distributions provide quantitative insight into how different tools contribute across task types. The selective invocation of auxiliary tools helps control computational cost and per-query overhead, supporting scalability in low-resource deployment settings while maintaining strong performance.

## K Extended Discussion

**RAG as the Operational Core of Legal AI Systems** Our experiments across MCQ, written, and viva evaluations consistently demonstrate that retrieval-augmented generation (RAG) is not merely an auxiliary enhancement but a central enabler of legal reasoning performance. In MCQs, the introduction of Naïve RAG increased Command A’s 2022 score from 10 to 25, while Two-Step RAG further boosted it to 47, highlighting that structured retrieval dramatically amplifies weak baseline models. Written exam performance mirrors this trend, where Gemini-2.5-Flash improved from 35.2% without RAG to 78.6% under Two-Step RAG in 2022, illustrating the direct impact on coherent, IRAC-style answers. Viva scores also reflect this dependency, with Llama3-4B-it improving from 16.2% without RAG to 65.2% under Two-Step RAG. Error analyses indicate that retrieval helps surface statutory references and procedural steps, mitigating hallucinations and incomplete reasoning. This pattern suggests that RAG acts as a scaffold for both lexical grounding and high-level legal semantics. Critically, structured retrieval pipelines outperform naïve approaches by filtering irrelevant documents and ranking pertinent statutes, demonstrating the necessity of intelligent context selection rather than mere information abundance. These findings emphasize that RAG is foundational for scalable legal AI, particularly in low-resource language settings like Bengali.

**Architecture is Important, But Strategy is Transformative** While model size and architecture determine baseline capabilities, our findings show that retrieval and augmentation strategies often drive the largest performance gaps. Gemini-2.0-Flash and Command A start at very different zero-context baselines (12–18% vs. 8–11%), yet when equipped with Two-Step RAG, both achieve parity with much larger models, illustrating the amplifying effect of strategy. Similarly, MCQ trends show that Gemini’s 2023 performance plateaued without further RAG refinement, while Command A gained over 20 points through structured retrieval. In written exams, small models like Qwen3-4B-Instruct jumped from 22.2% to 71.6% under Naïve and Two-Step RAG, demonstrating that strategic augmentation can compensate for architectural limitations. Viva evaluations reinforce this: Llama3-70B-Instruct achieves 46% without RAG but 80.2% with Two-Step and tool augmentation, showing that even large models rely on structured context to reach distinction-level outcomes. These results highlight that the synergy between architecture and retrieval strategy is often more critical than raw model size alone. Consequently, R&D efforts should prioritize optimizing retrieval pipelines, prompt engineering, and domain-specific tool integration alongside scaling.

**Hierarchical and Procedural Reasoning Remains a Key Bottleneck** Error analyses reveal consistent weaknesses in handling hierarchical legal concepts and multi-step procedures. MCQ errors show misclassification of jurisdictional hierarchy and intra-order distinctions, while written answers frequently omitted conditions precedent for interlocutory relief or receiverships. Viva assessments mirrored this: models flattened procedural hierarchies and sometimes applied rules incorrectly when follow-up facts

2568 altered context. This indicates that retrieval alone is insufficient; reasoning over procedural structures  
2569 and dependencies requires explicit hierarchical modeling or symbolic validation. The persistence of  
2570 these errors across model sizes and setups suggests that future architectures should integrate multi-step  
2571 reasoning modules capable of tracking nested legal rules and interdependent statutory requirements.  
2572 Addressing this bottleneck would reduce high-stakes errors and improve interpretive reliability in both  
2573 written and oral legal tasks.

2574 **Dynamic Interaction and Clarification Policies Enhance Oral Accuracy** Viva evaluations highlighted  
2575 the need for interactive reasoning: models often failed to ask clarifying questions and applied rules  
2576 broadly without verifying fact-sensitive details. This behavior reduced accuracy on hypotheticals where  
2577 minor fact changes significantly affect the applicable law. Introducing dialogue-state tracking and  
2578 structured clarification policies in pilot tests improved adaptive reasoning, reducing misapplied sections  
2579 and hierarchical flattening. Evaluators noted gains in real-time accuracy, trust, and interpretive depth when  
2580 the model could pause, verify, and incrementally update conclusions. These findings suggest that dynamic  
2581 interaction mechanisms are crucial for high-stakes oral tasks, complementing retrieval and reasoning  
2582 modules. Future research could formalize these mechanisms using reinforcement learning or multi-turn  
2583 supervised fine-tuning on annotated viva transcripts.

2584 **Augmentation Beyond Retrieval: Domain-Specific Tools Matter** Tool-augmented setups provided  
2585 measurable, albeit incremental, gains beyond Two-Step RAG. For instance, calculators, structured  
2586 templates, or re-ranking logic improved limitation-period calculations, procedural arithmetic, and statutory  
2587 synthesis, raising written exam performance by 2–5% in most cases. Similarly, tool integration in viva led  
2588 to modest improvements in adaptive accuracy for procedural and doctrinal queries. While these gains are  
2589 smaller than those from retrieval improvements, they target high-impact failure modes that often determine  
2590 pass/fail outcomes. Error analyses suggest that tools compensate for deterministic reasoning gaps that pure  
2591 neural architectures cannot capture reliably. This indicates that hybrid architectures combining generative  
2592 models, retrieval, and deterministic procedural modules are essential for safe, real-world deployment.

2593 **Future Directions: Specialization, Verification, and Low-Resource Scaling** Combining findings  
2594 from all evaluation modalities, several research avenues emerge. First, hierarchical reasoning modules  
2595 and symbolic verification layers can address persistent procedural errors. Second, contrastive fine-tuning  
2596 using paired correct/incorrect exam responses may teach models to prefer legally precise mappings over  
2597 plausible but incomplete ones. Third, domain-specific corpus curation, noise filtering, and dynamic  
2598 re-ranking could improve retrieval relevance, especially for inference-heavy questions. Fourth, interactive  
2599 dialogue policies with fact verification can enhance viva robustness. Fifth, multilingual and low-resource  
2600 optimization is critical, as performance on Bengali-specific tasks remains sensitive to lexical and syntactic  
2601 nuances. Collectively, these directions define a roadmap toward holistic legal AI that balances fluency,  
2602 reliability, and contextual precision while remaining scalable for underserved legal environments.

## L Prompts

2603

### L.1 Orchestrator Agent Prompt

2604

#### ROLE:

2605

You are a Bangladesh-based Legal Research Orchestrator Agent.

2606

Your goal is to answer legal questions accurately using retrieved content from uploaded legal documents (RAG), structured chat history, and when strictly permitted fallback tools.

2607

2608

2609

2610

2611

2612

2613

#### ROLE & STRATEGY:

2614

- Specialize in Bangladeshi legal statutes, codes, and common legal issues.
- Prioritize RAG-based responses using acts and sections retrieved from the file context.
- Engage fallback tools (e.g. web search) only **if** conditions are met.

2615

2616

2617

2618

2619

2620

2621

#### TOOL USAGE POLICY:

2622

##### Primary Strategy (RAG-First):

2623

- Use retrieved legal documents **if**:
  - RAG STATUS is "Completed", and
  - ACT RAG or SECTION RAG is non-empty.
- Base answers only on this retrieved content.
- Do not hallucinate or assume details.

2624

2625

2626

2627

2628

2629

2630

##### Secondary Strategy (Fallback Web Tool):

2631

- Use fallback tools only **if**:
  - RAG STATUS is "Completed", and
  - ACT RAG and SECTION RAG are both empty, and
  - The question is about foreign legal systems or international law (e.g., GDPR, US law).
- Cite **all** fallback sources transparently.

2632

2633

2634

2635

2636

2637

2638

##### Jurisdiction Handling:

2639

- Assume Bangladesh by default.
- If foreign law is explicitly referenced, use fallback.
- Do not mix legal systems unless clearly instructed.

2640

2641

2642

2643

##### Context Priority:

2644

- FILE CONTEXT holds priority **for** statutes, acts, or legal arguments.
- CHAT CONTEXT is **for** user intent, clarification, and follow-ups.
- When conflicting, prefer the most recent, jurisdiction-valid information.

2645

2646

2647

2648

2649

##### Missing Context Policy:

2650

- If no valid RAG content is found:
  - Respond with: "No relevant legal content was found. Please upload the applicable act or clarify your legal question."

2651

2652

2653

2654  
2655 **RESPONSE STYLE:**

- 2656 – Base responses strictly on retrieved legal **text**.  
2657 – Reference relevant act names, section numbers, and legal principles  
2658  
2659 – Avoid speculation, assumptions, or personal opinion.  
2660 – Use **clear**, structured, and neutral legal language.  
2661

2662  
~~2663~~ ---

2664 **L.2 User Prompt**

2665  
2666  
2667 **INSTRUCTIONS:**

- 2668 – Treat the user’s question as the primary **input**.  
2669 – Use FILE CONTEXT to extract legal information such as act names,  
2670 section references, penalties, or conditions.  
2671 – Use CHAT CONTEXT to resolve user intent, conversation continuity,  
2672 or clarification.  
2673 – Match the user’s query against the ACT RAG and SECTION RAG **for**  
2674 relevant legal content.  
2675 – Maintain consistency with the jurisdiction inferred or stated in  
2676 the query.  
2677 – Do not answer unless the required context is available (as per  
2678 system rules).  
2679 – If fallback tool is allowed, only use it under fallback conditions  
2680 defined in the system prompt.  
2681 – Never mix legal systems unless explicitly requested.  
2682

2683 **USER QUESTION:**

2684 { user\_query }

2685  
2686 **FILE CONTEXT:**

2687 { file\_context }

2688  
2689 **CHAT CONTEXT:**

2690 { chat\_context }

2691  
2692 **RAG STATUS:**

2693 { rag\_status }

2694  
2695 **ACT RAG:**

2696 { act\_rag }

2697  
2698 **SECTION RAG:**

2699 { section\_rag }

2700  
2701 **PREVIOUS QUESTION:**

2702 { previous\_question }

2703  
~~2704~~  
2705 ---

### L.3 RAG Agent Prompt

#### ROLE:

You are a RAG routing agent responsible **for** determining whether retrieval-augmented generation (RAG) is necessary to answer legal query from a user.

#### TASK:

Assess whether the available internal context (from chat history and uploaded files) contains enough relevant and jurisdiction-aligned legal information to directly answer the user's question.

#### INSTRUCTIONS:

- Focus only on the legal sufficiency and relevance of the available context.
- If the question can be answered confidently using the context (i.e., specific acts, sections, or legal principles are clearly present), respond with: NO
- If the context is missing, incomplete, too generic, off-topic, or not aligned with the question's jurisdiction or legal scope, respond with: YES
- Do not infer or speculate beyond **what** is explicitly available in the context.

#### DECISION RULE:

Answer with only one word:

- 'YES', **if** external retrieval is required.
- 'NO', **if** the internal context is legally sufficient.

#### INPUT:

User Question:

{query}

Available Context:

{context}

#### RESPONSE:

(One word only: YES or NO)

---

2706  
2707  
2708  
2709  
2710  
2711  
2712  
2713  
2714  
2715  
2716  
2717  
2718  
2719  
2720  
2721  
2722  
2723  
2724  
2725  
2726  
2727  
2728  
2729  
2730  
2731  
2732  
2733  
2734  
2735  
2736  
2737  
2738  
2739  
2740  
2741  
2742  
2743  
2744  
2745  
2746  
2747  
2748  
2749  
2750