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

Azmine Toushik Wasi [1 2 3] **Wahid Faisal** [1 2 3] **Mst Rafia Islam** [1 2 4] **Md Rizwan Parvez** [5]

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]Computational Intelligence and Operations Laboratory (CIOL) [2]Cohere Labs Community [3]Shahjalal University of Science and Technology, Bangladesh [4]Independent University, Bangladesh [5]Qatar Computing Research Institute (QCRI), Qatar. Correspondence to: Azmine Toushik Wasi <azmine32@student.sust.edu>.

Accepted to the 6th Muslims in ML (MusIML) Workshop, co-located with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

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 & 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 insertion. 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 & 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 procedures 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) 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 shown in Figure 1.

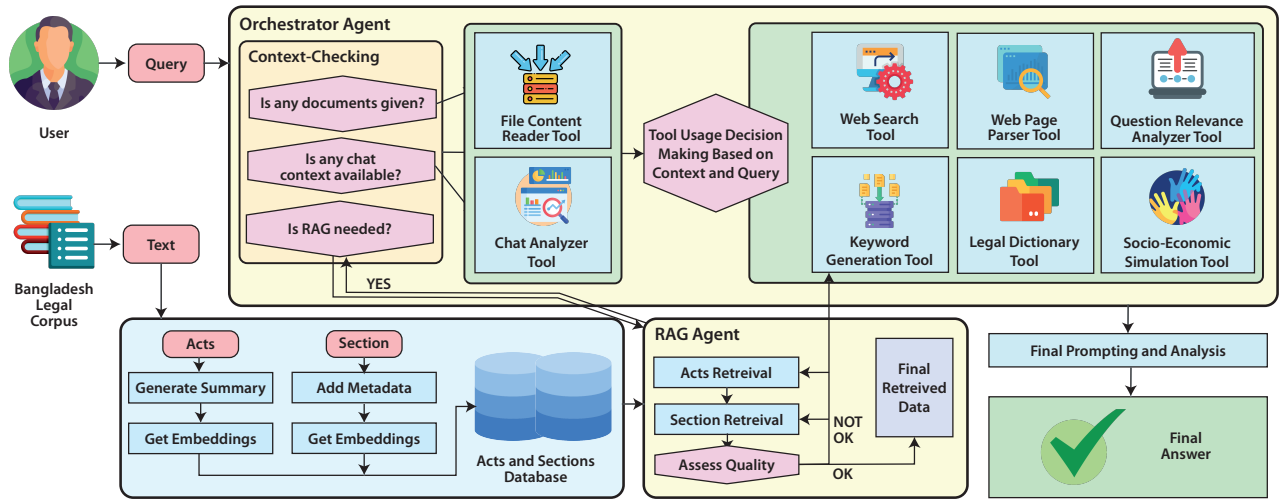

Figure 1. System Architecture and Workflow of our Multilingual Legal Assistant Agent for Bangladesh

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 relevance and jurisdictional appropriateness of retrieved legal materials. This separation of roles enhances system modularity and simplifies maintenance, debugging, and extensibility.

Both agents operate within a `LangGraph`-based state machine, enabling persistent memory across interactions. This design supports complex, multi-turn legal consultations where reasoning must evolve. The agentic structure also enables conditional execution, allowing the system to adapt its workflow dynamically without compromising performance or legal integrity.

3.1.3. TOOLS

Our system incorporates a suite of specialized tools, each tailored to handle specific preprocessing or auxiliary tasks critical to the overall workflow. The **File Content Reader** processes uploaded documents in *.pptx*, *.docx*, and *.pdf* formats, employing temporary file handling alongside format-

specific parsers to extract clean, structured text for downstream use. During the RAG phase, the **Keyword Generator** leverages a large language model (LLM) to produce 5–10 semantically meaningful search terms, with a regular expression fallback mechanism to ensure robustness in low-context scenarios. To facilitate external information access, a DuckDuckGo-based **Web Search** tool retrieves top-ranked results, including titles, URLs, and content snippets. When deeper inspection is required, the **Web Page Parser** employs `BeautifulSoup` to extract up to 5,000 characters of visible page content, filtering out scripts and styling artifacts. For internal coherence, the **Question Relevance Analyzer** utilizes embeddings to detect semantic relationships between current and prior queries, thereby enhancing continuity in multi-turn interactions. **Chat Analyzer** analyzes previous chat to get context in chat mode. Legal interpretation in Bengali is supported by a custom **Legal Dictionary**, designed to explain colonial-era and Farsi-influenced terms prevalent in Bangladeshi legal texts. This tool enables plain-language explanations and improves the model's ability to interpret complex expressions. Furthermore, a **Socio-Economic Simulation** module enables exploration of how socio-demographic variables interact with legal access, supporting both diagnostic use cases and long-term policy planning.

3.1.4. LLM INTEGRATION AND PROMPTS

At the core of our system's understanding layer is a multilingual LLM. In our experiments, we evaluate a broad spectrum of models spanning both proprietary and open-source families, including OpenAI's *GPT-4o* (OpenAI et al., 2024), Google's *gemini-2.0-flash* (proprietary), *gemini-2.5-flash* (proprietary) (Team et al., 2025a), *gemma-3-4b-it*, *gemma-3-12b-it*, *gemma-3-27b-it* (Team et al., 2025b), Meta's *llama3.2-1b-instruct*, *llama3.2-3b-instruct*, *llama3.1-8b-instruct*, *llama3.1-70b-instruct* (Grattafiori et al., 2024), Cohere's *command-a-03-2025* (Cohere et al., 2025), Qwen's *qwen3-4b-instruct-2507*, and *qwen3-30b-a3b-instruct-2507* (Yang et al., 2025). All models were evaluated across system modules for legal drafting, question answering, and prompt compliance.

For retrieval, we employ Cohere's *embed-multilingual-light-v3.0*, which supports semantic similarity in both Bengali and English, crucial for processing bilingual legal content in Bangladesh. Prompt engineering ensures legal accuracy and jurisdictional relevance. Prompts are constructed using prior conversation turns, uploaded documents, and the RAG state. All prompts are detailed in Appendix L.

3.2. Adaptive Agent Workflow

Our system operates through two primary working modes: *situational* and *decisional*. In the **situational** pathway, if a user uploads a document, it is routed through the Document Analyzer, which extracts and prepares relevant legal content for downstream use. Alternatively, when prior user interactions are available, the Chat Analyzer identifies related past queries to provide contextual grounding for the current request. In the **decisional** pathway, the system assesses whether retrieval-augmented generation (RAG) is necessary. When internal context, such as documents or chat history, is insufficient or absent, we trigger the RAG workflow. This involves generating semantic keywords, retrieving relevant vectors from pre-indexed legal corpora (e.g., Bangladeshi statutes and sections), and composing a contextually grounded response. This flexible routing mechanism enables the agent to adapt intelligently based on the type of input, historical interactions, and the complexity of the query, ensuring both responsiveness and legal relevance.

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-

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**.

Model	Year: 2022				Year: 2023			
	W/o RAG	Naive RAG	2-Step RAG	Tools	W/o RAG	Naive RAG	2-Step RAG	Tools
Random Choice	25.00	25.20	25.40	25.20	25.60	25.40	25.20	25.00
Proprietary Models								
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	**30.20**	**68.80**	**75.60**	**77.00**	**32.40**	**69.20**	**76.40**	**77.00**
Open-Source Models: Small								
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	**14.20**	**28.40**	44.60	**49.80**	**16.20**	**29.40**	45.40	**52.40**
Command-A-8B	8.20	25.20	**47.00**	47.40	11.20	23.40	**49.20**	50.20
Llama3.1-8B-Instruct	15.20	18.40	30.20	32.40	16.20	18.60	33.40	35.20
Open-Source Models: Large								
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	**34.20**	**50.40**	**65.60**	**70.80**	**36.20**	**52.40**	**67.20**	**72.40**
Llama3.1-70B-Instruct	24.20	30.40	40.20	42.40	26.20	32.20	44.40	46.20

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 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

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	Year: 2022				Year: 2023			
	W/o RAG	Naive RAG	2-Step RAG	Tools	W/o RAG	Naive RAG	2-Step RAG	Tools
Proprietary Models								
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	**35.20**	**70.40**	**78.60**	**81.00**	**36.20**	**71.20**	**79.40**	**81.80**
Open-Source Models: Small								
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	**71.60**	68.00	26.20	60.40	64.20	70.40
Command-A-8B	25.00	**60.20**	71.00	**74.40**	25.20	**73.00**	**74.20**	**76.00**
Llama3.1-8B-Instruct	**28.20**	34.40	46.20	50.00	**30.20**	66.00	68.20	71.20
Open-Source Models: Large								
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	**45.00**	**62.20**	**75.00**	**79.80**	**47.20**	**65.00**	**77.20**	**80.20**

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**.

Model	W/o RAG	N. RAG	2-S RAG	Tools
Proprietary Models				
Gemini-2.0 Flash	32.80	60.80	74.40	76.60
Gemini-2.5 Flash	**36.20**	**70.40**	**79.20**	**81.00**
Open-Source Models: Small				
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	**56.40**	67.20	70.20
Command A	**27.80**	55.60	**70.40**	**71.20**
Llama3.1-8B-Instruct	24.20	52.20	66.40	67.20
Open-Source Models: Large				
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	**46.00**	**63.20**	**77.20**	**80.20**

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, con-

sistently 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. Discussions

Integrated View of Retrieval, Reasoning, and Tools. The findings collectively demonstrate that effective legal AI systems require a tightly integrated combination of retrieval, reasoning, and tool-based augmentation rather than reliance on any single component. Retrieval-augmented generation serves as the primary mechanism for grounding responses in statutory knowledge, but its effectiveness depends heavily on structured pipelines that prioritize relevance and procedural context. At the same time, reasoning limitations—particularly in hierarchical and multi-step legal processes—highlight the need for architectures that go beyond surface-level text generation. Tool augmentation further complements these capabilities by addressing deterministic gaps such as procedural calculations and structured synthesis. Together, these components form a hybrid system in which neural generation, symbolic validation, and retrieval operate in tandem. This integrated perspective underscores that legal AI performance is fundamentally a systems-level problem rather than a model-centric one.

From Static Answers to Interactive Legal Intelligence. A key insight from viva evaluations is the importance of dynamic interaction in achieving reliable legal reasoning. Unlike static written responses, oral examination settings expose weaknesses in adaptability, fact verification, and incremental reasoning. The absence of clarification mechanisms often leads to misapplication of statutes when critical facts are ambiguous or evolving. Incorporating dialogue-state tracking and structured clarification policies enables the system to iteratively refine its understanding, improving both accuracy and user trust. This shift transforms the agent from a passive answer generator into an interactive reasoning system capable of engaging in legally meaningful dialogue. As such, future legal AI systems must be designed with multi-turn interaction as a core capability rather than an auxiliary feature.

Toward Robust and Equitable Legal AI Deployment. The broader implication of this work lies in the pathway toward deploying legal AI in real-world, resource-constrained environments. Persistent challenges such as procedural errors, incomplete reasoning, and sensitivity to linguistic variation highlight the importance of robustness and inclusivity in system design. Advances in contrastive fine-tuning, structured retrieval, and post-generation verification can significantly enhance reliability, particularly for high-stakes legal applications. Simultaneously, multilingual and dialect-aware optimization is essential to ensure accessibility in contexts like Bangladesh, where legal discourse spans multiple linguistic registers. User interface safeguards, including transparency mechanisms and risk indicators, further contribute to responsible deployment. Collectively, these directions point toward legal AI systems that are not only technically competent but also contextually aware, trustworthy, and equitable in practice.

9. Conclusion

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.

Ethical Considerations

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 ICML's code of ethics regarding research integrity, fairness, and respect for contributors.

Limitations

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.

Ethical Considerations

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.

Potential Risks

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.

Acknowledgements

We sincerely thank Cohere for providing computing credits through the Cohere Labs Research Grant, which were instrumental in developing Knowledge Explorer. We are deeply grateful to Cohere Labs for their continued mentorship and for organizing Aya Expedition 2.0, which offered an excellent environment for experimentation, collaboration, and constructive feedback.

We are deeply grateful to Julia Kreutzer (*Senior Research Scientist at Cohere Labs*) for her generous mentorship and guidance. Her insight, encouragement, and constructive feedback were invaluable throughout the conception, development, and refinement of MINA, shaping both its technical direction and research depth. We also thank Alejandro Rodríguez Salamanca, Marzieh Fadaee, Sara Hooker, Madeline Smith, and Brittawnya Prince for their insightful discussions, encouragement, and support during the expedition. We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey.

We would also like to express our sincere gratitude to the members of the expert panel for their valuable time, guidance, and thoughtful feedback throughout the development of this work. We are especially grateful to Prof. Dr. Muhammad Mahbubur Rahman, Prof. Dr. Liton Chandra Biswas, Mr. Md Al Ifran Hossain Mollah, Mr. Washik Muhammod Istiaz Ezaz, and Ms. Afroza Bilkis for their insightful comments and domain expertise, which helped refine the system design and analysis. We note that these individuals were not part of the formal evaluation process; all quantitative assessments were conducted independently by the designated evaluator group described in Section F.1.

References

Ahmed, A. Challenges with access to justice in bangladesh. *Centre for Governance Studies*, 2021.

Akter, F. Legal aid for ensuring access to justice in bangladesh: A paradox? *Asian Journal of Law and Society*, 4(1):257–275, jan 2017. ISSN 2052-9023. URL http://dx.doi.org/10.1017/ALS.2016.60.

Ariai, F. and Demartini, G. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges, 2025. URL https://arxiv.org/abs/2410.21306.

Asjad, T. Use of bangla in judicial system. February 2023. URL https://thefinancialexpress.com.bd/views/opinions/use-of-bangla-in-judicial-system. Accessed: 2025-06-22.

Cohere, T., :, Aakanksha, Ahmadian, A., Ahmed, M., Alammar, J., Alizadeh, M., Alnumay, Y., Althammer, S., Arkhangorodsky, A., Aryabumi, V., Aumiller, D., Avalos, R., Aviv, Z., Bae, S., Baji, S., Barbet, A., Bartolo, M., Bebensee, B., Beladia, N., Beller-Morales, W., Bérard, A., Berneshawi, A., Bialas, A., Blunsom, P., Bobkin, M., Bongale, A., Braun, S., Brunet, M., Cahyawijaya, S., Cairuz, D., Campos, J. A., Cao, C., Cao, K., Castagné, R., Cendrero, J., Currie, L. C., Chandak, Y., Chang, D., Chatziveroglou, G., Chen, H., Cheng, C., Chevalier, A., Chiu, J. T., Cho, E., Choi, E., Choi, E., Chung, T., Cirik, V., Cismaru, A., Clavier, P., Conklin, H., Crawhall-Stein, L., Crouse, D., Cruz-Salinas, A. F., Cyrus, B., D'souza, D., Dalla-Torre, H., Dang, J., Darling, W., Domingues, O. D., Dash, S., Debugne, A., Dehaze, T., Desai, S., Devassy, J., Dholakia, R., Duffy, K., Edalati, A., Eldeib, A., Elkady, A., Elsharkawy, S., Ergün, I., Ermis, B., Fadaee, M., Fan, B., Fayoux, L., Flet-Berliac, Y., Frosst, N., Gallé, M., Galuba, W., Garg, U., Geist, M., Azar, M. G., Gilsenan-McMahon, E., Goldfarb-Tarrant, S., Goldsack, T., Gomez, A., Gonzaga, V. M., Govindarajan, N., Govindassamy, M., Grinsztajn, N., Gritsch, N., Gu, P., Guo, S., Haefeli, K., Hajjar, R., Hawes, T., He, J., Hofstätter, S., Hong, S., Hooker, S., Hosking, T., Howe, S., Hu, E., Huang, R., Jain, H., Jain, R., Jakobi, N., Jenkins, M., Jordan, J., Joshi, D., Jung, J., Kalyanpur, T., Kamalakara, S. R., Kedrzycki, J., Keskin, G., Kim, E., Kim, J., Ko, W.-Y., Kocmi, T., Kozakov, M., Kryściński, W., Jain, A. K., Teru, K. K., Land, S., Lasby, M., Lasche, O., Lee, J., Lewis, P., Li, J., Li, J., Lin, H., Locatelli, A., Luong, K., Ma, R., Mach, L., Machado, M., Magbitang, J., Lopez, B. M., Mann, A., Marchisio, K., Markham, O., Matton, A., McKinney, A., McLoughlin, D., Mokry, J., Morisot, A., Moulder, A., Moynehan, H., Mozes, M., Muppalla, V., Murakhovska, L., Nagarajan, H., Nandula, A., Nasir, H., Nehra, S., Netto-Rosen, J., Ohashi, D., Owers-Bardsley, J., Ozuzu, J., Padilla, D., Park, G., Passaglia, S., Pekmez, J., Penstone, L., Piktus, A., Ploeg, C., Poulton, A., Qi, Y., Raghvendra, S., Ramos, M., Ranjan, E., Richemond, P., Robert-Michon, C., Rodriguez, A., Roy, S., Ruder, S., Ruis, L., Rust, L., Sachan, A., Salamanca, A., Saravanakumar, K. K., Satyakam, I., Sebag, A. S., Sen, P., Sepehri, S., Seshadri, P., Shen, Y., Sherborne, T., Shi, S. S., Shivaprasad, S., Shmyhlo, V., Shrinivason, A., Shteinbuk, I., Shukayev, A., Simard, M., Snyder, E., Spataru, A., Spooner, V., Starostina, T., Strub, F., Su, Y., Sun, J., Talupuru, D., Tarassov, E., Tommasone, E., Tracey, J., Trend, B., Tumer, E., Üstün, A., Venkitesh, B., Venuto, D., Verga, P., Voisin, M., Wang, A., Wang, D., Wang, S., Wen, E., White, N., Willman, J., Winkels, M., Xia, C., Xie, J., Xu, M., Yang, B., Yi-Chern, T., Zhang, I., Zhao, Z., and Zhao, Z. Command a: An enterprise-ready large language model, 2025. URL https://arxiv.org/abs/2504.00698.

Cui, J., Ning, M., Li, Z., Chen, B., Yan, Y., Li, H., Ling, B., Tian, Y., and Yuan, L. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model, 2024. URL https://arxiv.org/abs/2306.16092.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, January 2024. ISSN 1946-5319. doi: 10.1093/jla/laae003. URL http://dx.doi.org/10.1093/jla/laae003.

de Kinderen, S. and Winter, K. *Towards Taming Large Language Models with Prompt Templates for Legal GRL Modeling*, pp. 213–228. Springer Nature Switzerland, 2024. ISBN 9783031610073. doi: 10.1007/978-3-031-61007-3_17.

Frankenreiter, J. and Nyarko, J. *Natural Language Processing in Legal Tech*, pp. 70–90. Cambridge University Press, 2023.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat,

L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Homoki, P. and Ződi, Z. Large language models and their possible uses in law. *Hungarian Journal of Legal Studies*, 64(3):435–455, June 2024. ISSN 2560-1067. doi: 10.1556/2052.2023.00475. URL http://dx.doi.org/10.1556/2052.2023.00475.

Huang, J.-H., Yang, C.-C., Shen, Y., Pacces, A. M., and Kanoulas, E. Optimizing numerical estimation and operational efficiency in the legal domain through large language models, 2024. URL https://arxiv.org/abs/2407.19041.

Hussain, A. S. and Thomas, A. Large language models

for judicial entity extraction: A comparative study, 2024. URL https://arxiv.org/abs/2407.05786.

Islam, M. J., Suzuki, M., and Mazumder, N. Promoting access to justice in bangladesh: Towards a hybrid justice model. *International Journal of Law, Crime and Justice*, 77:100655, jun 2024. ISSN 1756-0616. doi: 10.1016/j.ijlcj.2024.100655. URL http://dx.doi.org/10.1016/j.ijlcj.2024.100655.

Jaan, A. R. What makes access to justice elusive in bangladesh? May 2023. URL https://blogs.lse.ac.uk/southasia/2023/05/01/what-makes-access-to-justice-elusive-in-bangladesh/. Accessed: 2025-06-22.

Jiang, C. and Yang, X. Agents on the bench: Large language model based multi agent framework for trustworthy digital justice, 2024. URL https://arxiv.org/abs/2412.18697.

Jiang, H., Zhang, X., Mahari, R., Kessler, D., Ma, E., August, T., Li, I., Pentland, A., Kim, Y., Roy, D., and Kabbara, J. Leveraging large language models for learning complex legal concepts through storytelling. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7194–7219, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.388. URL https://aclanthology.org/2024.acl-long.388.

Khalilullah, M. I. Justice on hold: Addressing bangladesh's judicial backlog. January 2025. URL https://www.tbsnews.net/thoughts/justice-hold-addressing-bangladeshs-judicial-backlog-1039686. Accessed: 2025-06-23.

Knock, S. A., Faruk, D. M. O., Siddiqua, R., and Akter, A. The practice of rule of law in criminal justice system: A criminological study of bangladesh. *International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences*, 12(4), August 2024. ISSN 2349-7300. doi: 10.37082/ijirmps.v12.i4.230737. URL http://dx.doi.org/10.37082/ijirmps.v12.i4.230737.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. Large language models in law: A survey, 2023. URL https://arxiv.org/abs/2312.03718.

Li, H., Chen, J., Yang, J., Ai, Q., Jia, W., Liu, Y., Lin, K., Wu, Y., Yuan, G., Hu, Y., Wang, W., Liu, Y., and Huang, M. Legalagentbench: Evaluating llm agents in legal domain, 2024. URL https://arxiv.org/abs/2412.17259.

Liang, G. and Tong, Q. Llm-powered ai agent systems and their applications in industry, 2025. URL https://arxiv.org/abs/2505.16120.

Louis, A., Van Dijck, G., and Spanakis, G. Interpretable long-form legal question answering with retrieval-augmented large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (20):22266–22275, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i20.30232. URL http://dx.doi.org/10.1609/aaai.v38i20.30232.

Mamalis, M., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., and Tarabanis, K. A large language model based legal assistant for governance applications, May 2024a. URL osf.io/94fmq.

Mamalis, M. E., Kalampokis, E., Fitsilis, F., Theodorakopoulos, G., and Tarabanis, K. *A Large Language Model Agent Based Legal Assistant for Governance Applications*, pp. 286–301. Springer Nature Switzerland, 2024b. ISBN 9783031702747. doi: 10.1007/978-3-031-70274-7_18. URL http://dx.doi.org/10.1007/978-3-031-70274-7_18.

Martin, L., Whitehouse, N., Yiu, S., Catterson, L., and Perera, R. Better call gpt, comparing large language models against lawyers, 2024. URL https://arxiv.org/abs/2401.16212.

Mik, E. Caveat lector: Large language models in legal practice, 2024. URL https://arxiv.org/abs/2403.09163.

Mizan, A. S. The law of language and the language of the law: A sociolegal appraisal of colonial legal language in bangladesh. *Dhaka University Law Journal*, 32(2): 119–134, 2021. doi: 10.3329/dulj.v32i2.57961.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette,

C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H. P., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M. O. T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Peppin, A., Kreutzer, J., Sebag, A. S., Marchisio, K., Ermis, B., Dang, J., Cahyawijaya, S., Singh, S., Goldfarb-Tarrant, S., Aryabumi, V., Aakanksha, Ko, W.-Y., Üstün, A., Gallé, M., Fadaee, M., and Hooker, S. The multilingual divide and its impact on global ai safety, 2025. URL https://arxiv.org/abs/2505.21344.

Qin, W. and Sun, Z. Exploring the nexus of large language models and legal systems: A short survey, 2024. URL https://arxiv.org/abs/2404.00990.

Raj, M. A. S. Assessing the legal framework and socioeconomic impacts of compensation for wrongfully convicted and imprisoned persons in bangladesh: Challenges and policy recommendations. *Cognizance Journal of Multidisciplinary Studies*, 4(6):456–466, June 2024. ISSN 0976-7797. doi: 10.47760/cognizance.2024.v04i06.028. URL http://dx.doi.org/10.47760/cognizance.2024.v04i06.028.

Safdie, L. Ai and legal aid: A generational opportunity for access to justice. February 2025. URL https://www.thomsonreuters.com/en-us/posts/ai-in-courts/ai-legal-aid-generational-opportunity/. Accessed: 2025-06-23.

Sun, J., Dai, C., Luo, Z., Chang, Y., and Li, Y. Lawluo: A chinese law firm co-run by llm agents, 2024. URL https://arxiv.org/abs/2407.16252.

Tahura, U. S. and Alam, S. Legal aid, systemic barriers, and access to justice in bangladesh. April 2025. URL https://www.thedailystar.net/law-our-rights/news/legal-aid-systemic-barriers-and-access-justice-ba Accessed: 2025-06-23.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Güra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Ágoston Weisz, Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M. A., Baeuml, M., Chen, Z., Shafey, L. E., Zhang, Y., Sercinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rrustemi, A., Clay, N., Crone, P., Kocisky, T., Zhao, J., Perz, B., Yu, D., Howard, H., Bloniarz, A., Rae, J. W., Lu, H., Sifre, L., Maggioni, M., Alcober, F., Garrette, D., Barnes, M., Thakoor, S., Austin, J., Barth-Maron, G., Wong, W., Joshi, R., Chaabouni, R., Fatiha, D., Ahuja, A., Tomar, G. S., Senter, E., Chadwick, M., Kornakov, I., Attaluri, N., Iturrate, I., Liu, R., Li, Y., Cogan, S., Chen, J., Jia, C., Gu, C., Zhang, Q., Grimstad, J., Hartman, A. J., Garcia, X., Pillai, T. S., Devlin, J., Laskin, M., de Las Casas, D., Valter, D., Tao, C., Blanco, L., Badia, A. P., Reitter, D., Chen, M., Brennan, J., Rivera, C., Brin, S., Iqbal, S., Surita, G., Labanowski, J., Rao, A., Winkler, S., Parisotto, E., Gu, Y., Olszewska, K., Addanki, R., Miech, A., Louis, A., Teplyashin, D., Brown, G., Catt, E., Balaguer, J., Xiang, J., Wang, P., Ashwood, Z., Briukhov, A., Webson, A., Ganapathy, S., Sanghavi, S., Kannan, A., Chang, M.-W., Stjerngren, A., Djolonga, J., Sun, Y., Bapna, A., Aitchison, M., Pejman, P., Michalewski, H., Yu, T., Wang, C., Love, J., Ahn, J., Bloxwich, D., Han, K., Humphreys, P., Sellam, T., Bradbury, J., Godbole, V., Samangooei, S., Damoc, B., Kaskasoli, A., Arnold, S. M. R., Vasudevan, V., Agrawal, S., Riesa, J., Lepikhin, D., Tanburn, R., Srinivasan, S., Lim, H., Hodkinson, S., Shyam, P., Ferret, J., Hand, S., Garg, A., Paine, T. L., Li, J., Li, Y., Giang, M., Neitz, A., Abbas, Z., York, S., Reid, M., Cole, E., Chowdhery, A., Das, D., Rogozińska, D., Nikolaev, V., Sprechmann, P., Nado, Z., Zilka, L., Prost, F., He, L., Monteiro, M., Mishra, G., Welty, C., Newlan, J., Jia, D., Allamanis, M., Hu, C. H., de Liedekerke, R., Gilmer, J., Saroufim, C., Rijhwani, S., Hou, S., Shrivastava, D., Baddepudi, A., Goldin, A., Ozturel, A., Cassirer, A., Xu, Y., Sohn, D., Sachan, D., Amplayo, R. K., Swanson, C., Petrova, D., Narayan, S., Guez, A., Brahma, S., Landon, J., Patel, M., Zhao, R., Villela, K., Wang, L., Jia, W., Rahtz, M., Giménez, M., Yeung, L., Keeling, J., Georgiev, P., Mincu, D., Wu, B., Haykal, S., Saputro, R., Vodrahalli, K., Qin, J., Cankara, Z., Sharma, A., Fernando, N., Hawkins, W., Neyshabur, B., Kim, S., Hutter, A., Agrawal, P., Castro-Ros, A., van den Driessche, G., Wang, T., Yang, F., yiin Chang, S., Komarek, P., McIlroy, R., Lučić, M., Zhang, G., Farhan, W., Sharman, M., Natsev, P., Michel, P., Bansal, Y., Qiao, S., Cao, K., Shakeri, S., Butterfield, C., Chung, J., Rubenstein, P. K., Agrawal, S., Mensch, A., Soparkar, K., Lenc, K., Chung, T., Pope, A., Maggiore, L., Kay, J., Jhakra, P., Wang, S., Maynez, J., Phuong, M., Tobin, T., Tacchetti, A., Trebacz, M., Robinson, K., Katariya, Y., Riedel, S., Bailey, P., Xiao, K., Ghelani, N., Aroyo, L., Slone, A., Houlsby, N., Xiong, X., Yang, Z., Gribovskaya, E., Adler, J., Wirth, M., Lee, L., Li, M., Kagohara, T., Pavagadhi, J., Bridgers, S., Bortsova, A., Ghemawat, S., Ahmed, Z., Liu, T., Powell, R., Bolina, V., Iinuma, M., Zablotskaia, P., Besley, J., Chung, D.-W., Dozat, T., Comanescu, R., Si, X., Greer, J., Su, G., Polacek, M., Kaufman, R. L., Tokumine, S., Hu, H., Buchatskaya, E., Miao, Y., Elhawaty, M., Siddhant, A., Tomasev, N., Xing, J., Greer, C., Miller, H., Ashraf, S., Roy, A., Zhang, Z., Ma, A., Filos, A., Besta, M., Blevins, R., Klimenko, T., Yeh, C.-K., Changpinyo, S., Mu, J., Chang, O., Pajarskas, M., Muir, C., Cohen, V., Lan, C. L., Haridasan, K., Marathe, A., Hansen, S., Douglas, S., Samuel, R., Wang, M., Austin, S., Lan, C., Jiang, J., Chiu, J., Lorenzo, J. A., Sjösund, L. L., Cevey, S., Gleicher, Z., Avrahami, T., Boral, A., Srinivasan, H., Selo, V., May, R., Aisopos, K., Hussenot, L., Soares, L. B., Baumli, K., Chang, M. B., Recasens, A., Caine, B., Pritzel, A., Pavetic, F., Pardo, F., Gergely, A., Frye, J., Ramasesh, V., Horgan, D., Badola, K., Kassner, N., Roy, S., Dyer, E., Campos, V. C., Tomala, A., Tang, Y., Badawy, D. E., White, E., Mustafa, B., Lang, O., Jindal, A., Vikram, S., Gong, Z., Caelles, S., Hemsley, R., Thornton, G., Feng, F., Stokowiec, W., Zheng, C., Thacker, P., Çağlar Ünlü, Zhang, Z., Saleh, M., Svensson, J., Bileschi, M., Patil, P., Anand, A., Ring, R., Tsihlas, K., Vezer, A., Selvi, M., Shevlane, T., Rodriguez, M., Kwiatkowski, T., Daruki, S., Rong, K., Dafoe, A., FitzGerald, N., Gu-Lemberg, K., Khan, M., Hendricks, L. A., Pellat, M., Feinberg, V., Cobon-Kerr, J., Sainath, T., Rauh, M., Hashemi, S. H., Ives, R., Hasson, Y., Noland, E., Cao, Y., Byrd, N., Hou, L., Wang, Q., Sottiaux, T., Paganini, M., Lespiau, J.-B., Moufarek, A., Hassan, S., Shivakumar, K., van Amersfoort, J., Mandhane, A., Joshi, P., Goyal, A., Tung, M., Brock, A., Sheahan, H., Misra, V., Li, C., Rakićević, N., Dehghani, M., Liu, F., Mittal, S., Oh, J., Noury, S., Sezener, E., Huot, F., Lamm, M., Cao, N. D., Chen, C., Mudgal, S., Stella, R., Brooks, K., Vasudevan, G., Liu, C., Chain, M., Melinkeri, N., Cohen, A., Wang, V., Seymore,

K., Zubkov, S., Goel, R., Yue, S., Krishnakumaran, S., Albert, B., Hurley, N., Sano, M., Mohananey, A., Joughin, J., Filonov, E., Kępa, T., Eldawy, Y., Lim, J., Rishi, R., Badiezadegan, S., Bos, T., Chang, J., Jain, S., Padmanabhan, S. G. S., Puttagunta, S., Krishna, K., Baker, L., Kalb, N., Bedapudi, V., Kurzrok, A., Lei, S., Yu, A., Litvin, O., Zhou, X., Wu, Z., Sobell, S., Siciliano, A., Papir, A., Neale, R., Bragagnolo, J., Toor, T., Chen, T., Anklin, V., Wang, F., Feng, R., Gholami, M., Ling, K., Liu, L., Walter, J., Moghaddam, H., Kishore, A., Adamek, J., Mercado, T., Mallinson, J., Wandekar, S., Cagle, S., Ofek, E., Garrido, G., Lombriser, C., Mukha, M., Sun, B., Mohammad, H. R., Matak, J., Qian, Y., Peswani, V., Janus, P., Yuan, Q., Schelin, L., David, O., Garg, A., He, Y., Duzhyi, O., Älgmyr, A., Lottaz, T., Li, Q., Yadav, V., Xu, L., Chinien, A., Shivanna, R., Chuklin, A., Li, J., Spadine, C., Wolfe, T., Mohamed, K., Das, S., Dai, Z., He, K., von Dincklage, D., Upadhyay, S., Maurya, A., Chi, L., Krause, S., Salama, K., Rabinovitch, P. G., M, P. K. R., Selvan, A., Dektiarev, M., Ghiasi, G., Guven, E., Gupta, H., Liu, B., Sharma, D., Shtacher, I. H., Paul, S., Akerlund, O., Aubet, F.-X., Huang, T., Zhu, C., Zhu, E., Teixeira, E., Fritze, M., Bertolini, F., Marinescu, L.-E., Bölle, M., Paulus, D., Gupta, K., Latkar, T., Chang, M., Sanders, J., Wilson, R., Wu, X., Tan, Y.-X., Thiet, L. N., Doshi, T., Lall, S., Mishra, S., Chen, W., Luong, T., Benjamin, S., Lee, J., Andrejczuk, E., Rabiej, D., Ranjan, V., Styrc, K., Yin, P., Simon, J., Harriott, M. R., Bansal, M., Robsky, A., Bacon, G., Greene, D., Mirylenka, D., Zhou, C., Sarvana, O., Goyal, A., Andermatt, S., Siegler, P., Horn, B., Israel, A., Pongetti, F., Chen, C.-W. L., Selvatici, M., Silva, P., Wang, K., Tolins, J., Guu, K., Yogev, R., Cai, X., Agostini, A., Shah, M., Nguyen, H., Donnaile, N. O., Pereira, S., Friso, L., Stambler, A., Kurzrok, A., Kuang, C., Romanikhin, Y., Geller, M., Yan, Z., Jang, K., Lee, C.-C., Fica, W., Malmi, E., Tan, Q., Banica, D., Balle, D., Pham, R., Huang, Y., Avram, D., Shi, H., Singh, J., Hidey, C., Ahuja, N., Saxena, P., Dooley, D., Potharaju, S. P., O'Neill, E., Gokulchandran, A., Foley, R., Zhao, K., Dusenberry, M., Liu, Y., Mehta, P., Kotikalapudi, R., Safranek-Shrader, C., Goodman, A., Kessinger, J., Globen, E., Kolhar, P., Gorgolewski, C., Ibrahim, A., Song, Y., Eichenbaum, A., Brovelli, T., Potluri, S., Lahoti, P., Baetu, C., Ghorbani, A., Chen, C., Crawford, A., Pal, S., Sridhar, M., Gurita, P., Mujika, A., Petrovski, I., Cedoz, P.-L., Li, C., Chen, S., Santo, N. D., Goyal, S., Punjabi, J., Kappaganthu, K., Kwak, C., LV, P., Velury, S., Choudhury, H., Hall, J., Shah, P., Figueira, R., Thomas, M., Lu, M., Zhou, T., Kumar, C., Jurdi, T., Chikkerur, S., Ma, Y., Yu, A., Kwak, S., Ähdel, V., Rajayogam, S., Choma, T., Liu, F., Barua, A., Ji, C., Park, J. H., Hellendoorn, V., Bailey, A., Bilal, T., Zhou, H., Khatir, M., Sutton, C., Rzadkowski, W., Macintosh, F., Vij, R., Shagin, K., Medina, P., Liang, C., Zhou, J., Shah, P., Bi, Y., Dankovics, A., Banga, S., Lehmann, S., Bredesen, M., Lin, Z., Hoffmann, J. E., Lai, J., Chung, R., Yang, K., Balani, N., Bražinskas, A., Sozanschi, A., Hayes, M., Alcalde, H. F., Makarov, P., Chen, W., Stella, A., Snijders, L., Mandl, M., Kärrman, A., Nowak, P., Wu, X., Dyck, A., Vaidyanathan, K., R, R., Mallet, J., Rudominer, M., Johnston, E., Mittal, S., Udathu, A., Christensen, J., Verma, V., Irving, Z., Santucci, A., Elsayed, G., Davoodi, E., Georgiev, M., Tenney, I., Hua, N., Cideron, G., Leurent, E., Alnahlawi, M., Georgescu, I., Wei, N., Zheng, I., Scandinaro, D., Jiang, H., Snoek, J., Sundararajan, M., Wang, X., Ontiveros, Z., Karo, I., Cole, J., Rajashekhar, V., Tumeh, L., Ben-David, E., Jain, R., Uesato, J., Datta, R., Bunyan, O., Wu, S., Zhang, J., Stanczyk, P., Zhang, Y., Steiner, D., Naskar, S., Azzam, M., Johnson, M., Paszke, A., Chiu, C.-C., Elias, J. S., Mohiuddin, A., Muhammad, F., Miao, J., Lee, A., Vieillard, N., Park, J., Zhang, J., Stanway, J., Garmon, D., Karmarkar, A., Dong, Z., Lee, J., Kumar, A., Zhou, L., Evens, J., Isaac, W., Irving, G., Loper, E., Fink, M., Arkatkar, I., Chen, N., Shafran, I., Petrychenko, I., Chen, Z., Jia, J., Levskaya, A., Zhu, Z., Grabowski, P., Mao, Y., Magni, A., Yao, K., Snaider, J., Casagrande, N., Palmer, E., Suganthan, P., Castaño, A., Giannoumis, I., Kim, W., Rybiński, M., Sreevatsa, A., Prendki, J., Soergel, D., Goedeckemeyer, A., Gierke, W., Jafari, M., Gaba, M., Wiesner, J., Wright, D. G., Wei, Y., Vashisht, H., Kulizhskaya, Y., Hoover, J., Le, M., Li, L., Iwuanyanwu, C., Liu, L., Ramirez, K., Khorlin, A., Cui, A., LIN, T., Wu, M., Aguilar, R., Pallo, K., Chakladar, A., Perng, G., Abellan, E. A., Zhang, M., Dasgupta, I., Kushman, N., Penchev, I., Repina, A., Wu, X., van der Weide, T., Ponnapalli, P., Kaplan, C., Simsa, J., Li, S., Dousse, O., Yang, F., Piper, J., Ie, N., Pasumarthi, R., Lintz, N., Vijayakumar, A., Andor, D., Valenzuela, P., Lui, M., Paduraru, C., Peng, D., Lee, K., Zhang, S., Greene, S., Nguyen, D. D., Kurylowicz, P., Hardin, C., Dixon, L., Janzer, L., Choo, K., Feng, Z., Zhang, B., Singhal, A., Du, D., McKinnon, D., Antropova, N., Bolukbasi, T., Keller, O., Reid, D., Finchelstein, D., Raad, M. A., Crocker, R., Hawkins, P., Dadashi, R., Gaffney, C., Franko, K., Bulanova, A., Leblond, R., Chung, S., Askham, H., Cobo, L. C., Xu, K., Fischer, F., Xu, J., Sorokin, C., Alberti, C., Lin, C.-C., Evans, C., Dimitriev, A., Forbes, H., Banarse, D., Tung, Z., Omernick, M., Bishop, C., Sterneck, R., Jain, R., Xia, J., Amid, E., Piccinno, F., Wang, X., Banzal, P., Mankowitz, D. J., Polozov, A., Krakovna, V., Brown, S., Bateni, M., Duan, D., Firoiu, V., Thotakuri, M., Natan, T., Geist, M., tan Girgin, S., Li, H., Ye, J., Roval, O., Tojo, R., Kwong, M., Lee-Thorp, J., Yew, C., Sinopalnikov, D., Ramos, S., Mellor, J., Sharma, A., Wu, K., Miller, D., Sonnerat, N., Vnukov, D., Greig, R., Beattie, J., Caveness, E., Bai, L., Eisenschlos, J., Korchemniy, A., Tsai, T., Jasarevic, M., Kong, W., Dao, P., Zheng, Z., Liu, F., Yang, F., Zhu, R., Teh, T. H., Sanmiya,

J., Gladchenko, E., Trdin, N., Toyama, D., Rosen, E., Tavakkol, S., Xue, L., Elkind, C., Woodman, O., Carpenter, J., Papamakarios, G., Kemp, R., Kafle, S., Grunina, T., Sinha, R., Talbert, A., Wu, D., Owusu-Afriyie, D., Du, C., Thornton, C., Pont-Tuset, J., Narayana, P., Li, J., Fatehi, S., Wieting, J., Ajmeri, O., Uria, B., Ko, Y., Knight, L., Héliou, A., Niu, N., Gu, S., Pang, C., Li, Y., Levine, N., Stolovich, A., Santamaria-Fernandez, R., Goenka, S., Yustalim, W., Strudel, R., Elqursh, A., Deck, C., Lee, H., Li, Z., Levin, K., Hoffmann, R., Holtmann-Rice, D., Bachem, O., Arora, S., Koh, C., Yeganeh, S. H., Põder, S., Tariq, M., Sun, Y., Ionita, L., Seyedhosseini, M., Tafti, P., Liu, Z., Gulati, A., Liu, J., Ye, X., Chrzaszcz, B., Wang, L., Sethi, N., Li, T., Brown, B., Singh, S., Fan, W., Parisi, A., Stanton, J., Koverkathu, V., Choquette-Choo, C. A., Li, Y., Lu, T., Ittycheriah, A., Shroff, P., Varadarajan, M., Bahargam, S., Willoughby, R., Gaddy, D., Desjardins, G., Cornero, M., Robenek, B., Mittal, B., Albrecht, B., Shenoy, A., Moiseev, F., Jacobsson, H., Ghaffarkhah, A., Rivière, M., Walton, A., Crepy, C., Parrish, A., Zhou, Z., Farabet, C., Radebaugh, C., Srinivasan, P., van der Salm, C., Fidjeland, A., Scellato, S., Latorre-Chimoto, E., Klimczak-Plucińska, H., Bridson, D., de Cesare, D., Hudson, T., Mendolicchio, P., Walker, L., Morris, A., Mauger, M., Guseynov, A., Reid, A., Odoom, S., Loher, L., Cotruta, V., Yenugula, M., Grewe, D., Petrushkina, A., Duerig, T., Sanchez, A., Yadlowsky, S., Shen, A., Globerson, A., Webb, L., Dua, S., Li, D., Bhupatiraju, S., Hurt, D., Qureshi, H., Agarwal, A., Shani, T., Eyal, M., Khare, A., Belle, S. R., Wang, L., Tekur, C., Kale, M. S., Wei, J., Sang, R., Saeta, B., Liechty, T., Sun, Y., Zhao, Y., Lee, S., Nayak, P., Fritz, D., Vuyyuru, M. R., Aslanides, J., Vyas, N., Wicke, M., Ma, X., Eltyshev, E., Martin, N., Cate, H., Manyika, J., Amiri, K., Kim, Y., Xiong, X., Kang, K., Luisier, F., Tripuraneni, N., Madras, D., Guo, M., Waters, A., Wang, O., Ainslie, J., Baldridge, J., Zhang, H., Pruthi, G., Bauer, J., Yang, F., Mansour, R., Gelman, J., Xu, Y., Polovets, G., Liu, J., Cai, H., Chen, W., Sheng, X., Xue, E., Ozair, S., Angermueller, C., Li, X., Sinha, A., Wang, W., Wiesinger, J., Koukoumidis, E., Tian, Y., Iyer, A., Gurumurthy, M., Goldenson, M., Shah, P., Blake, M., Yu, H., Urbanowicz, A., Palomaki, J., Fernando, C., Durden, K., Mehta, H., Momchev, N., Rahimtoroghi, E., Georgaki, M., Raul, A., Ruder, S., Redshaw, M., Lee, J., Zhou, D., Jalan, K., Li, D., Hechtman, B., Schuh, P., Nasr, M., Milan, K., Mikulik, V., Franco, J., Green, T., Nguyen, N., Kelley, J., Mahendru, A., Hu, A., Howland, J., Vargas, B., Hui, J., Bansal, K., Rao, V., Ghiya, R., Wang, E., Ye, K., Sarr, J. M., Preston, M. M., Elish, M., Li, S., Kaku, A., Gupta, J., Pasupat, I., Juan, D.-C., Someswar, M., M., T., Chen, X., Amini, A., Fabrikant, A., Chu, E., Dong, X., Muthal, A., Buthpitiya, S., Jauhari, S., Hua, N., Khandelwal, U., Hitron, A., Ren, J., Rinaldi, L., Drath, S., Dabush, A., Jiang, N.-J., Godhia, H., Sachs, U., Chen, A., Fan, Y., Taitelbaum, H., Noga, H., Dai, Z., Wang, J., Liang, C., Hamer, J., Ferng, C.-S., Elkind, C., Atias, A., Lee, P., Listík, V., Carlen, M., van de Kerkhof, J., Pikus, M., Zaher, K., Müller, P., Zykova, S., Stefanec, R., Gatsko, V., Hirnschall, C., Sethi, A., Xu, X. F., Ahuja, C., Tsai, B., Stefanoiu, A., Feng, B., Dhandhania, K., Katyal, M., Gupta, A., Parulekar, A., Pitta, D., Zhao, J., Bhatia, V., Bhavnani, Y., Alhadlaq, O., Li, X., Danenberg, P., Tu, D., Pine, A., Filippova, V., Ghosh, A., Limonchik, B., Urala, B., Lanka, C. K., Clive, D., Sun, Y., Li, E., Wu, H., Hongtongsak, K., Li, I., Thakkar, K., Omarov, K., Majmundar, K., Alverson, M., Kucharski, M., Patel, M., Jain, M., Zabelin, M., Pelagatti, P., Kohli, R., Kumar, S., Kim, J., Sankar, S., Shah, V., Ramachandruni, L., Zeng, X., Bariach, B., Weidinger, L., Vu, T., Andreev, A., He, A., Hui, K., Kashem, S., Subramanya, A., Hsiao, S., Hassabis, D., Kavukcuoglu, K., Sadovsky, A., Le, Q., Strohman, T., Wu, Y., Petrov, S., Dean, J., and Vinyals, O. Gemini: A family of highly capable multimodal models, 2025a. URL https://arxiv.org/abs/2312.11805.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P.,

Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025b. URL https://arxiv.org/abs/2503.19786.

Wasi, A. T., Faisal, W., Islam, M. R., and Bappy, M. M. Exploring possibilities of ai-powered legal assistance in bangladesh through large language modeling, 2024. URL https://arxiv.org/abs/2410.17210.

Wehnert, S. Justifiable artificial intelligence: Engineering large language models for legal applications, 2023. URL https://arxiv.org/abs/2311.15716.

Yan, Q. Legal challenges of artificial intelligence in the field of criminal defense. *Lecture Notes in Education Psychology and Public Media*, 30(1):167–175, December 2023. ISSN 2753-7056. doi: 10.54254/2753-7048/30/20231629. URL http://dx.doi.org/10.54254/2753-7048/30/20231629.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yesmin, S. Alternative dispute resolution: A strategic solution to bangladesh's judicial backlog. May 2025. URL https://www.tbsnews.net/thoughts/alternative-dispute-resolution-strategic-solution-bangladeshs-judicial-backlog-1134421. Accessed: 2025-06-23.

Yong, Z.-X., Ermis, B., Fadaee, M., Bach, S. H., and Kreutzer, J. The state of multilingual llm safety research: From measuring the language gap to mitigating it, 2025. URL https://arxiv.org/abs/2505.24119.

Yue, S., Huang, T., Jia, Z., Wang, S., Liu, S., Song, Y., Huang, X., and Wei, Z. Multi-agent simulator drives language models for legal intensive interaction. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6537–6570, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.365. URL https://aclanthology.org/2025.findings-naacl.365/.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. How does NLP benefit legal system: A summary of legal artificial intelligence. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5218–5230, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.466. URL https://aclanthology.org/2020.acl-main.466.

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 & 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 & 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 & 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 & Yang, 2024). Broader surveys of LLM agents detail architectures integrating RAG, hierarchical planning, and safety mechanisms for domain-focused applications (Liang & 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 & Sun, 2024) and legal text comprehension studies (Martin et al., 2024; Wehnert, 2023; Homoki & Ződi, 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 & Ződi, 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.

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,

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 _____ পর্যায়ের আদালতে।"

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)

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 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.

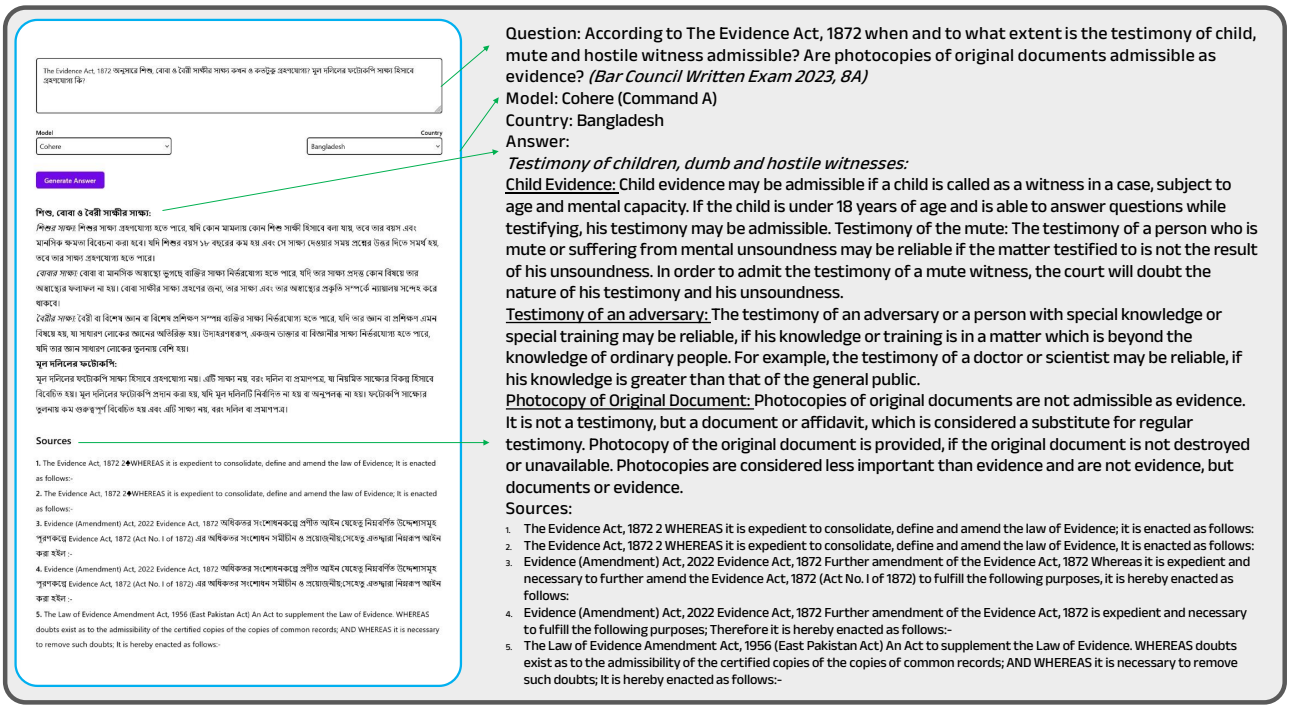

Figure 3. System Demonstration: UI and deployable system of MINA.

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.

E.4. Question Relevance Analyzer

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

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

E.5. Chat Analyzer

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

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

E.6. Legal Dictionary

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

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

E.7. Socio-Economic Simulation

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

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

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

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

F. Additional Information on Experiments and Evaluation

F.1. Evaluators

The evaluation of written and viva voce responses was conducted by law faculty members from leading Bangladeshi institutions, including the University of Dhaka and Independent University, Bangladesh (IUB), as well as practicing advocates enrolled in the Bangladesh Bar. This combination ensures both academic rigor and alignment with real-world legal practice.

For the viva voce examination, five evaluators independently scored each response, while three evaluators assessed the written examination. All evaluators routinely engage in legal instruction, examination, or courtroom practice, and therefore applied the same criteria used in official Bangladesh Bar Council examinations, including accuracy, clarity, contextual understanding, and legal reasoning. No additional task-specific training was provided, as evaluators were already familiar with established grading standards and marking practices.

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 further 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 κ (κ) between all pairs of evaluators. Each evaluator scored the 13 questions numerically along four dimensions: accuracy, clarity, contextual understanding, and legal reasoning. Cohen's κ is computed as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

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 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

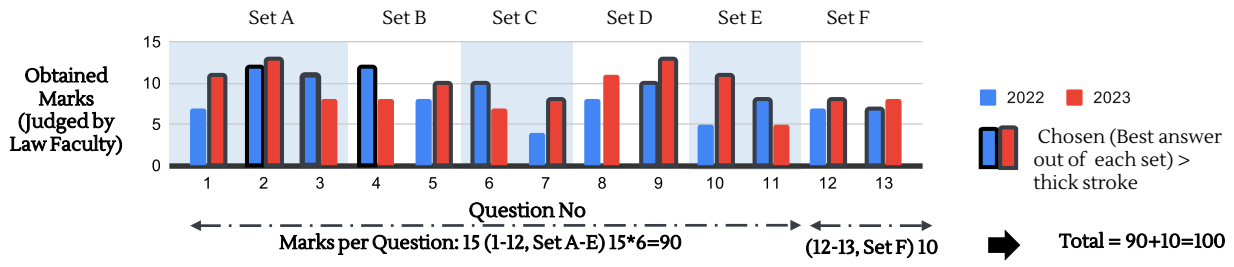

Figure 4. Breaking Down a Written Full Answer (Command-A, Two Step; Examiner 2)

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.

F.6. Addressing Potential Selection Bias in Written Assessment

A potential concern in the original evaluation protocol is the use of confidence-based filtering for written responses, which may introduce selection bias by preferentially including higher-confidence (and potentially higher-quality) outputs. To rigorously assess the extent of this effect, we conduct an additional evaluation on the full set of written questions, without any confidence-based selection.

Table 4 reports the performance of all models under this unfiltered evaluation setting, using five independent evaluators. Across all architectures and prompting strategies, we observe only a modest absolute performance decrease ranging from 1.0% to 3.8% compared to the originally reported confidence-filtered results. Importantly, the relative ranking of models and the comparative advantages of retrieval-augmented and tool-augmented settings remain consistent, indicating that the primary conclusions of the study are robust to the removal of selection effects.

To further quantify this effect, Table 5 presents the performance drop relative to confidence-based selection. The observed decreases are uniformly small across model families, scales, and evaluation configurations. Notably, even in lower-performing models, where one might expect greater sensitivity to filtering, the degradation remains limited and does not alter qualitative trends. Larger and more capable models (e.g., 30B–70B class) exhibit particularly stable behavior, with drops often below 2%, suggesting that confidence-based filtering does not disproportionately inflate their performance.

These findings indicate that while confidence-based selection introduces a measurable bias, its magnitude is insufficient to meaningfully affect the study's conclusions. The robustness of results under full-set evaluation strengthens the validity of our claims regarding the effectiveness of retrieval augmentation and tool integration for legal reasoning tasks.

Table 4. Evaluation of MINA on BD Bar Council Written Exams (All Questions, No Confidence Filtering). Scores are averaged across five evaluators.

Model	Year: 2022				Year: 2023			
	W/o RAG	Naive RAG	2-Step RAG	Tools	W/o RAG	Naive RAG	2-Step RAG	Tools
Proprietary Models								
GPT-4o	17.4	54.4	68.8	70.2	17.4	57.6	69.8	73.0
Gemini-2.0-Flash	17.4	58.2	67.6	69.0	15.6	63.6	73.6	74.4
Gemini-2.5-Flash	34.2	69.8	78.4	80.0	34.4	69.6	78.6	81.6
Open-Source Models: Small								
Llama3.2-1B-Instruct	3.8	4.2	5.2	6.6	3.8	4.0	5.2	6.4
Llama3.2-3B-Instruct	7.0	8.0	11.4	14.4	10.4	13.6	17.2	19.6
Gemma-3-4B-it	12.2	18.0	22.2	26.2	16.6	21.0	25.4	29.8
Qwen3-4B-Instruct-2507	20.2	55.0	70.2	67.6	23.4	58.0	61.8	68.0
Command-A-8B	21.4	56.6	68.0	72.2	21.6	70.4	71.6	74.0
Llama3.1-8B-Instruct	24.4	30.6	43.2	47.8	28.2	64.6	67.0	71.0
Open-Source Models: Large								
Gemma-3-12B-it	28.0	41.2	55.0	57.4	28.8	41.2	54.8	57.6
Gemma-3-27B-it	36.8	54.0	68.0	72.2	38.4	55.6	69.8	73.6
Qwen3-30B-A3B-Instruct-2507	40.4	59.0	73.2	77.4	41.0	59.6	73.8	77.6
Llama3.1-70B-Instruct	43.0	60.4	74.0	79.6	45.8	63.6	76.4	80.0

Table 5. Performance drop (%) compared to confidence-based selection.

Model	Year: 2022				Year: 2023			
	W/o RAG	Naive RAG	2-Step RAG	Tools	W/o RAG	Naive RAG	2-Step RAG	Tools
Proprietary Models								
GPT-4o	1.8	1.0	1.0	1.0	3.8	3.2	2.8	2.6
Gemini-2.0-Flash	1.0	0.0	1.0	1.0	2.6	1.8	1.2	0.8
Gemini-2.5-Flash	1.0	0.6	0.2	1.0	1.8	1.6	0.8	0.2
Open-Source Models: Small								
Llama3.2-1B-Instruct	2.4	1.8	1.0	0.4	3.2	3.0	2.0	1.6
Llama3.2-3B-Instruct	3.2	3.0	2.8	2.0	1.6	0.6	0.2	0.4
Gemma-3-4B-it	3.0	2.2	2.2	1.8	1.6	1.4	0.6	0.4
Qwen3-4B-Instruct-2507	2.0	1.4	1.4	0.4	2.8	2.4	2.4	2.4
Command-A-8B	3.6	3.6	3.0	2.2	3.6	2.6	2.6	2.0
Llama3.1-8B-Instruct	3.8	3.8	3.0	2.2	2.0	1.4	1.2	0.2
Open-Source Models: Large								
Gemma-3-12B-it	2.0	1.0	0.4	0.6	3.4	3.2	3.2	2.6
Gemma-3-27B-it	1.4	1.0	0.2	0.2	1.6	0.6	0.2	0.6
Qwen3-30B-A3B-Instruct-2507	1.8	1.4	0.8	0.8	3.0	2.6	2.6	1.8
Llama3.1-70B-Instruct	2.0	1.8	1.0	0.2	1.4	1.4	0.8	0.2

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

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

G.2. Written Exam

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

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

2022: 3A | Command A – 8B (with 2 Step RAG) | Mark 10 | Evaluator 1

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।
- কানুনের স্পষ্ট অধিকার (Clear Legal Right): যদি বিচারকের সামনে এটি প্রমাণ করা যায় যে বিচারবিভাগের কাছে স্পষ্ট কানুনি 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 অন্যান্য 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)

2022: 2A | Gemma-3-12B-it (with 2 Step RAG and Tools) | Mark 5 | Evaluator 3

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)

2022: 6A | Gemma-3-4B (with Naive RAG) | Mark 10 | Evaluator 2

Original Answer

Question: দণ্ডবিধি, ১৮৬০ এ সন্নিবেশিত সাধারণ ব্যতিক্রমসমূহ বর্ণনা করুন।
Answer: দণ্ডবিধি, ১৮৬০-এ কোনো সাধারণ ব্যতিক্রম নেই।

Translation

Question: Describe the general exceptions incorporated in the Penal Code, 1860.
Answer: The Penal Code, 1860 does not contain any general exceptions.

Evaluation

Mark: 0/10
Comment: Wrong Answer. The Penal Code does contain general exceptions (Sections 76–106), covering acts done under mistake of fact, accident, consent, self-defence, etc.

2022: 6B | Gemma-3-4B (with Naive RAG) | Mark 5 | Evaluator 2

Original Answer

Question: X, Y এর নিকট থেকে ২ লক্ষ টাকা ঋণ নেন 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 উল্লেখে কে পরামর্শ দিন।
Answer: 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, পরিস্থিতি অনুযায়ী প্রতারণা বা চুক্তি ভঙ্গের অভিযোগে দণ্ডবিধি, ১৮৬০ অনুসারে একটি ফৌজদারি মামলাও বিবেচনা করতে পারেন।
আইনি পরামর্শ:

- দেওয়ানি মামলা: Y সিভিল প্রসিডিউর কোড, ১৯০৮ অনুযায়ী, বিশেষ করে অর্ডার ৩৭ (সারসংক্ষেপ মামলা) এর অধীনে দ্রুত ঋণ পুনরুদ্ধারের জন্য মামলা দায়ের করতে পারেন।
- ফৌজদারি মামলা: যদি প্রতারণামূলক উদ্দেশ্য বা বিশ্বাসভঙ্গের প্রমাণ পাওয়া যায়, তবে Y দণ্ডবিধি, ১৮৬০ অনুসারে, বিশেষ করে ধারা ৪২০ (প্রতারণা এবং অসৎভাবে সম্পত্তি হস্তান্তর করানো) এর অধীনে মামলা দায়ের করতে পারেন

অধিক্ষেত্র (Jurisdiction): উভয় দেওয়ানি ও ফৌজদারি মামলা সেই আদালতে দায়ের করা উচিত, যার অধিক্ষেত্রের মধ্যে ঋণচুক্তি সম্পাদিত হয়েছে বা যেখানে X বসবাস করেন।
মন্তব্য: ব্যাখ্যাটি স্পষ্ট ও প্রাসঙ্গিক। তবে, চুক্তি ভঙ্গ একটি দেওয়ানি অন্যায়, এটি ফৌজদারি অপরাধ নয়। কেবলমাত্র যদি শুরু থেকেই প্রতারণার উদ্দেশ্য থাকে, তখনই দণ্ডবিধির ৪২০ ধারাটি প্রযোজ্য হতে পারে।

Translation

Question: (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..
Answer: 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, Y may consider filing a criminal case for breach of contract or fraud, depending on the circumstances, under the Penal Code, 1860.
Legal Advice:

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.
2. 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).

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.
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.

Evaluation

Mark: 3.5/5
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.

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

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.

G.2.3. POTENTIAL SOLUTIONS

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

H. Detailed Evaluator Response Analysis

H.1. Written Examination: detailed evaluation and patterns

We evaluated the written component using the fixed CQ question set described above, with each model response independently marked and annotated by five senior law faculty evaluators. Evaluators scored answers on accuracy, completeness, legal reasoning, statutory citation, and drafting quality, and provided line-by-line comments on content and structure. Across models, large retrieval-augmented agents produced responses that closely matched the expected exam structure: issue identification, statement of law, application to facts, and concise conclusion, facilitating straightforward marking.

These models reliably surfaced relevant statutory provisions and often included short drafting templates or sample prayers that examiners found practically useful. Smaller models, by contrast, frequently omitted statutory prerequisites or failed to synthesize multiple statutory sources, resulting in lower marks and critical comments. A recurring strength noted by evaluators was the agents' organization and clarity: many answers read like passable student scripts, employing the IRAC pattern that examiners reward. Evaluators praised bilingual fluency and plain-language paraphrases, which made complex doctrines teachable and testable in an exam setting. However, a common negative pattern was omission of procedural prerequisites, such as conditions precedent for appointing a receiver or prerequisites for interlocutory relief, which examiners considered exam-level faults rather than stylistic lapses. Procedural arithmetic tasks, including limitation-period calculations and adjournment timelines, were another failure mode; models sometimes miscounted days or ignored statutory tolling, prompting deductive marks. Inter-statutory synthesis proved challenging: when questions required aligning rules from the Specific Relief Act, CPC, and Evidence Act, agents sometimes produced plausible but incomplete mappings, omitting key interaction points. Annotative comments from evaluators highlighted hallucinated or misattributed case citations in a minority of responses, suggesting retrieval precision remains imperfect under time-constrained prompts. Evaluator feedback showed moderate inter-rater consistency in final marks but substantial variance in qualitative comments, reflecting different tolerance thresholds for omission versus stylistic choices. One examiner summarized typical written feedback as follows:

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

Others also commented:

> **E2**: I was impressed by how MINA maintained structured coherence even under conversational 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 findings highlight the system's potential as a scalable drafting and study aid while underscoring the necessity of deterministic, rule-aware modules for high-stakes legal outputs. All graded responses and anonymized evaluator comments are archived for future fine-tuning and error analysis to systematically close the remaining performance gaps.

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

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

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

> **E2:** The articulation of remedies under the Specific Relief Act was remarkably clear and concise. I appreciated

how MINA linked the relief sought to underlying equitable principles, a feature many human examinees neglect. It also demonstrated awareness of practical implications, noting how injunctions protect property interests before adjudication. Still, I found the explanation somewhat mechanical; the model stated the rule without fully exploring its exceptions or judicial discretion. In advanced answers, I expect an engagement with case illustration or critical reasoning, not merely doctrinal recall.

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

E4: 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.

E5: 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.

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.

H.3. Alignment with Bangladesh Bar Council Evaluation Standards

The written and viva evaluations were designed to closely replicate the implicit yet well-established grading criteria used in Bangladesh Bar Council examinations. Rather than introducing task-specific or NLP-centric metrics, we ground our evaluation framework in the same holistic dimensions employed by human legal examiners, thereby preserving ecological validity.

Accuracy. This dimension measures the factual and doctrinal correctness of a response. Evaluators assess whether the correct statutes, sections, and legal principles are identified, whether the elements of offenses are accurately stated, and whether procedural rules are applied without error. Incorrect section references, misclassification of offenses, or doctrinal misunderstandings result in direct score penalties, consistent with standard Bar Council marking practices.

Clarity. Clarity reflects how intelligibly and professionally the answer is presented. Evaluators examine logical organization, appropriate use of legal terminology, sentence structure, and overall readability. For written responses, this corresponds to the coherence and structural quality expected in human scripts. In viva settings, clarity additionally encompasses articulation, conciseness, and the ability to respond directly and unambiguously to examiner queries.

Contextual Understanding. This dimension evaluates whether the response correctly interprets the factual scenario and applies the law appropriately to the given context. Evaluators check for accurate issue identification, relevance of cited legal provisions, and avoidance of generic or boilerplate responses. In viva examinations, this further includes the ability to adapt reasoning when factual variations are introduced by the examiner, reflecting authentic oral examination dynamics.

Legal Reasoning. Legal reasoning captures the quality of analytical progression from facts to legal conclusions. Evaluators assess whether responses follow a recognizable IRAC-style structure, demonstrate causal linkage between facts and statutory elements, and provide justified conclusions rather than unsupported assertions. Responses are penalized when reasoning steps are incomplete, implicit, or entirely absent.

Scoring Procedure. Consistent with actual Bangladesh Bar Council evaluations, these dimensions are not scored independently. Instead, evaluators apply them holistically to assign a single composite score (e.g., out of 10, 15, or 100, depending on the examination format). This mirrors official grading practices, where scripts are evaluated as integrated legal performances rather than through decomposed rubrics. Such a design choice avoids overfitting to artificial metrics and ensures that the evaluation framework remains aligned with real-world legal assessment standards.

H.4. Qualitative Feedback Analysis

The qualitative feedback in Figure 8 illustrates that evaluators consistently assess model outputs using authentic legal examination standards, emphasizing doctrinal accuracy, structured reasoning (e.g., IRAC), and procedural completeness. High-performing systems are frequently described as "exam-ready" or "pass-level," indicating that their outputs align closely with human expectations in professional legal contexts. At the same time, recurring limitations, such as missing procedural details, reduced reasoning depth, or lack of adaptive responses in viva, highlight meaningful gaps that are not fully captured by quantitative scores. The inclusion of bilingual feedback further demonstrates that evaluators naturally operate across English and Bangla, reinforcing the ecological validity of the evaluation setup. Overall, these excerpts substantiate that the human-in-the-loop assessment captures both competence and nuanced failure modes in a realistic legal evaluation environment.

I. Extended Analysis

I.1. Detailed Findings from MCQ

Here we explore MCQ evalauiton results in more detail.

1. **Baseline Performance Highlights Model Capacity.** 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.

2. **Naïve RAG Provides Moderate Gains, Sensitive to Noise.** Introducing unfiltered retrieval boosts weaker models significantly but shows diminishing returns for top models. Command-A-8B increased from 8.2% $\rightarrow$ 25.2% in 2022 (+17 pts) and 11.2% $\rightarrow$ 23.4% in 2023 (+12.2 pts). Gemini-2.5-Flash improved from 30.2% $\rightarrow$ 68.8% (+38.6 pts) in 2022 and 32.4% $\rightarrow$ 69.2% (+36.8 pts) in 2023, indicating that strong models already leverage latent context. Some regression in 2023 (e.g., Gemini-2.0-Flash: 61.2% $\rightarrow$ 59.4%) reflects noise sensitivity and retrieval irrelevance. Small models like Llama3.2-1B see negligible gains. These trends indicate that Naïve RAG is beneficial but insufficient

Anonymized Faculty Quotes

- ✓ **Written exam, Evaluator 2, Gemini-2.5-Flash, 2-step RAG + tools.** "The answer follows an exam-ready IRAC structure with accurate statutory grounding. This is comparable to a solid pass-level human script."
- ✓ **Written exam, Evaluator 1, Qwen3-30B-A3B-Instruct-2507, tools.** "সুন্দর উত্তর। প্রাসঙ্গিক দণ্ডবিধির ধারাগুলো সঠিকভাবে শনাক্ত করতে পেরেছে এবং যুক্তিও স্পষ্ট। কিছু প্রক্রিয়াগত শর্ত বাদ পড়েছে, তবে সামগ্রিকভাবে আইনি বিশ্লেষণ সুসংগত।" ("Good Answer. Relevant Penal Code sections are correctly identified and the reasoning is clear. Some procedural requirements are missing, but the overall legal analysis is coherent.")
- ✓ **Written exam, Evaluator 3, Gemma-3-27B-it, tools**. "The bilingual explanation is clear and professionally phrased. This would score comfortably above the minimum threshold."
- ✓ **Viva, Evaluator 3, GPT-4o, 2-step RAG + tools.** "প্রশ্নোত্তরের সময় আত্মবিশ্বাস বজায় রেখেছে এবং ফলো-আপ প্রশ্নে সরাসরি প্রাসঙ্গিক ধারা উল্লেখ করেছে, যেটা আমি ভাবিনি।" ("The model maintained composure during questioning and cited relevant sections directly in follow-up questions, which was quite unexpected to me")
- ✓ **Viva, Evaluator 5, Gemini-2.0-Flash, tools.** "Clear articulation and correct section references. Lacked adaptive questioning, but responses were legally sound."
- ✓ **Viva, Evaluator 2, Llama3.1-70B-Instruct, tools.** "ডকট্রিনাল জ্ঞান ভালো এবং ব্যাখ্যা স্পষ্ট। মাঝে মাঝে দ্বিধা দেখা গেছে, তবে তা ভাইবা পরীক্ষায় খুব স্বাভাবিক। কনফিউজিং কোশ্চেনেও ভালো ফলোয়াপ ছিল। ভালোভাবে পাস করার মতোই রেজাল্ট" ("Good doctrinal knowledge and clear explanations. Some hesitation was observed, but it is normal for viva. Good follow-up in confusing questions. Overall, performance was definitely pass-worthy.")
- ✓ **Written exam, Evaluator 4, Command-A-8B, 2-step RAG.** "Correct statute identified, but procedural prerequisites are incomplete. Reasoning depth is limited."
- ✓ **Written exam, Evaluator 1, Qwen3-4B-Instruct-2507, naïve RAG.** "উত্তরটা বেশ সাবলীল, তবে একটা উদ্ধৃত ধারা ভুল ছিল। এই কারণে নম্বর কাটা যেত।" ("The answer is linguistically fluent, but one cited provision is wrong. This would result in a loss of marks.")

Figure 8. Anonymized evaluator feedback excerpts illustrating qualitative assessment of model responses across written and viva settings. Comments highlight strengths such as IRAC-structured reasoning and statutory accuracy, as well as limitations including missing procedural details and reduced adaptability. The bilingual nature of feedback reflects real-world legal evaluation practices in Bangladesh.

for high-stakes MCQs; structured filtering and relevance prioritization are critical to avoid noisy context misguiding weaker models.

3. **Two-Step RAG as a Game-Changer, Especially for Mid-Tier Models.** Filtering and reranking retrieved content yields the largest performance improvements. Command-A-8B jumps from 25.2% $\rightarrow$ 47.0% in 2022 and 23.4% $\rightarrow$ 49.2% in 2023. Gemma-3-12B-it improves 35.2% $\rightarrow$ 48.4% (2022) and 36.2% $\rightarrow$ 52.4% (2023). Even top-tier Qwen3-30B-A3B-Instruct-2507 increases from 50.4% $\rightarrow$ 65.6% (2022) and 52.4% $\rightarrow$ 67.2% (2023). Gains stem from reduced retrieval noise and prioritization of highly relevant statutes and precedents. Mid-tier models benefit disproportionately, as structured retrieval amplifies latent knowledge otherwise inaccessible. The pattern underscores that high-quality context is more impactful than sheer model size for exam performance.

4. **Diminishing Returns from Additional Tools.** Incorporating calculators, advanced prompt chaining, or re-ranking logic provides only marginal gains beyond Two-Step RAG. For instance, Qwen3-30B-A3B-Instruct-2507 increases 65.6% $\rightarrow$ 70.8% in 2022 and 67.2% $\rightarrow$ 72.4% in 2023. Similar trends appear for Command-A-8B and Gemini-3-27B-it. Once relevant context is available, auxiliary tools primarily assist procedural or arithmetic tasks, while deeper reasoning and intra-statutory synthesis remain bottlenecks. This plateau suggests that further gains require model-level improvements, not just tool stacking.

5. **Cross-Year Dynamics Reflect Exam Complexity and Model Adaptation.** From 2022 to 2023, weaker models (e.g., Command-A-8B) show steady Two-Step RAG gains (47.0% $\rightarrow$ 49.2%), while top models plateau (Gemini-2.5-Flash 75.6% $\rightarrow$ 76.4%). Naïve RAG slightly declines, implying more inference-heavy or ambiguous questions in 2023. Exam-specific reasoning, such as multi-step statutory synthesis and intra-order distinctions, remains challenging across models. Future improvement hinges on reasoning depth, retrieval precision, and contextual integration, rather than size or additional tools alone. The data illustrates an interplay between model architecture, retrieval strategy, and exam design shaping performance evolution.

I.2. Detailed Findings from Written Exam

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

Comparative Effectiveness: Naïve vs. Structured Retrieval Strategies Naïve RAG shows moderate but inconsistent improvements, particularly sensitive to irrelevant or noisy documents. For instance, Qwen3-4B-Instruct-2507 jumped from 22.2% → 56.4% in 2022 but achieved only 26.2% → 60.4% in 2023, reflecting retrieval sensitivity to question phrasing. Two-Step RAG consistently outperforms Naïve RAG, especially for mid-tier open-source models: Command-A-8B rises from 60.2% → 71.0% in 2022 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 $\rightarrow$ 70.4, Gemma-3-4B-it 50.4 $\rightarrow$ 65.2, and Gemini-2.5-Flash 70.4 $\rightarrow$ 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 $\rightarrow$ 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 $\rightarrow$ 81.0, Qwen3-30B-A3B-Instruct 75.2 $\rightarrow$ 79.4, and Command A 70.4 $\rightarrow$ 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 $\rightarrow$ 71.2 across all setups, while Llama3.1-70B-Instruct progresses 46.0 $\rightarrow$ 80.2. Small models such as Llama3.2-1B remain near floor (6.2 $\rightarrow$ 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 $\rightarrow$ 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.

I.4. Socio-Economic Simulation Module

The Socio-Economic Simulation module is designed to model how non-doctrinal factors influence real-world legal outcomes in Bangladesh, where access to justice is often shaped by income, occupation, literacy, geography, and social capital in addition to statutory law. Unlike the core retrieval-augmented generation (RAG) pipeline, which operates over codified legal sources, this module functions as a *diagnostic overlay* that contextualizes legal analysis within structural inequalities, without modifying the underlying statutory conclusions.

Worked Example (Bar Council Exam 2023, Q4(b)). **Scenario (abridged):** X assaults Y; Y, enraged, reacts violently. Z exploits Y's anger and provides Y with a knife, intending that Y kill X. Y subsequently kills X. Police charge Y and Z under Sections 109, 34, 323, and 326 of the Penal Code, 1860, and a *naraji* petition is to be prepared.

Simulation Inputs.

- **Accused Y:** Rural agricultural worker; monthly income $\approx$ 12,000 BDT; low literacy; limited access to legal counsel.
- **Accused Z:** Urban businessman; higher income; established social and professional networks.
- **Victim X:** Member of a marginalized community.

Simulation Outputs.

- **Pre-trial detention risk:** The module flags a higher likelihood of prolonged pre-trial detention for Y due to limited financial capacity to secure bail, reduced awareness of procedural remedies, and weaker access to early legal representation.
- **Bail and representation disparity:** Z is predicted to obtain anticipatory bail or legal counsel more rapidly, despite comparable statutory exposure, reflecting disparities driven by income and social capital rather than doctrinal differences.
- **Procedural delay exposure:** Y is identified as more vulnerable to repeated adjournments and delayed hearings, increasing custodial duration and pressure to accept unfavorable outcomes.
- **Socio-legal bias indicators:** The system highlights how rural location, low literacy, and occupational status may influence policing behavior, remand practices, and courtroom dynamics, even when statutory provisions are formally applied equally.

Interpretation and Role. Importantly, the module does not alter the legal analysis of liability under the Penal Code. Instead, it surfaces likely divergences between formal legal equality and lived legal outcomes, enabling users to understand why similarly situated defendants may experience different trajectories within the justice system. Accordingly, the module is best understood as a *socio-legal diagnostic and policy analysis component*, rather than a predictive sentencing or prescriptive advisory system.

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

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 6. Average tool-call frequency per query across exam stages.

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

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

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

K.1. Future Directions

While MINA demonstrates strong performance in structured legal reasoning and evaluation settings, several important directions remain for advancing both its technical robustness and real-world applicability.

Hierarchical and Symbolic Reasoning Integration. Future systems should incorporate structured legal knowledge bases and hierarchical reasoning frameworks to better capture procedural dependencies across court levels. In particular, a low-latency symbolic validation layer could enforce procedural constraints in real time, mitigating persistent errors in multi-step legal reasoning.

Deterministic Procedural Calculators. The integration of deterministic procedural calculators is essential for accurately handling timeline arithmetic. This would directly address observed failure modes in limitation-period calculations, including incorrect day counting and omission of statutory tolling considerations.

Interactive Clarification Policies. To improve robustness in viva-style interactions, the system should adopt interactive clarification policies. Specifically, when key facts are ambiguous, the agent should generate targeted follow-up questions to ensure accurate fact grounding before proceeding with legal analysis.

Dialogue-State Tracking. A dedicated dialogue-state tracking mechanism is needed to maintain and verify factual consistency across conversational turns. Such a system would enable the agent to detect contradictions, reference prior statements, and incrementally update legal conclusions in multi-turn reasoning scenarios.

Contrastive Fine-Tuning on Exam Data. Performance can be further enhanced through contrastive fine-tuning on annotated bilingual exam-style datasets. By training on paired correct and incorrect responses, the model can learn to prioritize precise statutory mappings, enforce procedural preconditions, and avoid superficially plausible but incomplete reasoning.

Post-Generation Verification Layer. A lightweight post-generation verification module should be introduced to validate cited statutes and detect missing procedural prerequisites. This layer would act as a safeguard against exam-level omissions while preserving the fluency of generated responses.

Structured Retrieval Fusion. Future work should extend the retrieval pipeline with structured fusion mechanisms that explicitly enforce statutory dependencies. Improvements in domain-specific corpus curation, noise filtering, and dynamic re-ranking will further enhance retrieval relevance, particularly for inference-heavy legal queries.

User Interface Safeguards. For real-world deployment, evaluators emphasized the importance of UI-level safeguards. These include providing provenance links for statutory citations, displaying confidence scores for key assertions, and incorporating an "examiner notes" panel that highlights unresolved risks or missing considerations.

Multilingual and Dialectal Expansion. Finally, continued optimization for low-resource and multilingual settings remains critical. Future research should systematically evaluate performance across Bengali dialects and register variations, with particular attention to sensitivity in legal interpretation arising from subtle lexical and syntactic differences.

Synthesis and Research Outlook. Combining insights across all evaluation modalities, several unifying research directions emerge. First, the integration of hierarchical reasoning modules with symbolic verification layers offers a principled solution to persistent procedural errors. Second, contrastive fine-tuning on paired correct and incorrect exam-style responses can encourage models to prefer legally precise statutory mappings over superficially plausible but incomplete answers. Third, advances in structured retrieval, including domain-specific corpus curation, noise filtering, and dynamic re-ranking, are essential for improving performance on inference-heavy queries. Fourth, interactive dialogue policies coupled with fact verification and dialogue-state tracking can substantially enhance robustness in viva-style settings. Finally, continued emphasis on multilingual and low-resource optimization is critical, as performance on Bengali legal tasks remains highly sensitive to lexical and syntactic variation. Collectively, these directions outline a roadmap toward legal AI systems that balance fluency, reliability, and contextual precision, while remaining scalable and equitable for underserved legal environments.

L. Prompts

L.1. Orchestrator Agent Prompt

```
ROLE:
You are a Bangladesh-based Legal Research Orchestrator Agent.

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.

ROLE & STRATEGY:

- 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.

TOOL USAGE POLICY:

Primary Strategy (RAG-First):
- 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.

Secondary Strategy (Fallback Web Tool):
- 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.

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

Context Priority:
- 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.

Missing Context Policy:
- 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."

RESPONSE STYLE:
- Base responses strictly on retrieved legal text.
- Reference relevant act names, section numbers, and legal principles.
- Avoid speculation, assumptions, or personal opinion.
- Use clear, structured, and neutral legal language.

,
```

L.2. User Prompt

```
INSTRUCTIONS:
- Treat the user's question as the primary input.
- Use FILE CONTEXT to extract legal information such as act names, section
    references, penalties, or conditions.
- Use CHAT CONTEXT to resolve user intent, conversation continuity, or
    clarification.
- Match the user's query against the ACT RAG and SECTION RAG for relevant legal
    content.
- Maintain consistency with the jurisdiction inferred or stated in the query.
- Do not answer unless the required context is available (as per system rules).
- If fallback tool is allowed, only use it under fallback conditions defined in
    the system prompt.
- Never mix legal systems unless explicitly requested.

USER QUESTION:
{user_query}

FILE CONTEXT:
{file_context}

CHAT CONTEXT:
{chat_context}

RAG STATUS:
{rag_status}
```

```
ACT RAG:
{act_rag}

SECTION RAG:
{section_rag}

PREVIOUS QUESTION:
{previous_question}

,
```

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)

,
```