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

Azmine Toushik Wasi[1,2,3*], Wahid Faisal[1,2,3*], Mst Rafia Islam[1,2,4]

[1]Computational Intelligence and Operations Laboratory [2]Cohere Labs Community
[3]Shahjalal University of Science and Technology [4]Independent University

[*]Equal Contribution

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 Exams, MINA scored 75–80% in Preliminary MCQs, Written, and simulated Viva Voce exams, matching or surpassing average human performance and demonstrating clarity, contextual understanding, and sound legal reasoning. Even under a conservative upper bound, MINA operates at just 0.12–0.61% of typical legal consultation costs in Bangladesh, yielding a 99.4–99.9% cost reduction relative to human-provided services. 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 case study on building domain-specific, low-resource systems and addressing challenges of multilingual adaptation, efficiency, and sustainable public-service AI deployment.

📅 **Date**: November 05, 2025
💻 **Code**: GitHub

✉️ **Correspondence**: Azmine Toushik Wasi (azmine32@student.sust.edu)

Computational Intelligence and Operations Lab

1. Introduction

Access to affordable legal advice remains a major challenge for economically disadvantaged groups in Bangladesh, largely due to complex legal language, opaque procedures, and high costs (Knock et al., 2024, Raj, 2024, Islam et al., 2024, Akter, 2017, Wasi et al., 2024). This contributes to a severe access to justice crisis that disproportionately affects marginalized populations. The judiciary is burdened by an enormous backlog of 3.7 to 4.4 million cases and operates with only 2,100 judges (Yesmin, 2025, Ahmed, 2021), translating to one judge per 90,000 people and resulting in delays of 10 to 60 years (Jaan, 2023). Civil disputes often drag on for decades, while criminal cases like the Chittagong Arms Haul have remained unresolved for over 15 years (Knock et al., 2024, Khalilullah, 2025). Outdated procedural laws enable frequent adjournments and exploitation by lawyers, further slowing case resolution. Meanwhile, unregulated lawyer fees make legal representation unaffordable for most (Jaan, 2023), and public legal aid faces constraints from limited funding, strict eligibility, and engagement (Tahura and Alam, 2025).

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

Motivated by the urgent need to improve legal access for Bangladesh's low-income population, who face significant barriers due to complex legal language, long procedures, and high costs, we developed Mina, a multilingual LLM-based legal assistant tailored to the Bangladeshi legal context. Unlike prior systems focused on English and Western legal frameworks, our assistant integrates Bengali and English support using multilingual embeddings within a Retrieval-Augmented Generation (RAG) framework that combines retrieval, reasoning, drafting, translation, and citation 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 and Nyarko, 2023, Zhong et al., 2020), with transformer models like BERT and GPT enhancing contextual understanding for more accurate insights (Martin et al., 2024, Jiang et al., 2024). Large Language Models (LLMs) extend this potential by simplifying complex legal language, translating statutes and 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

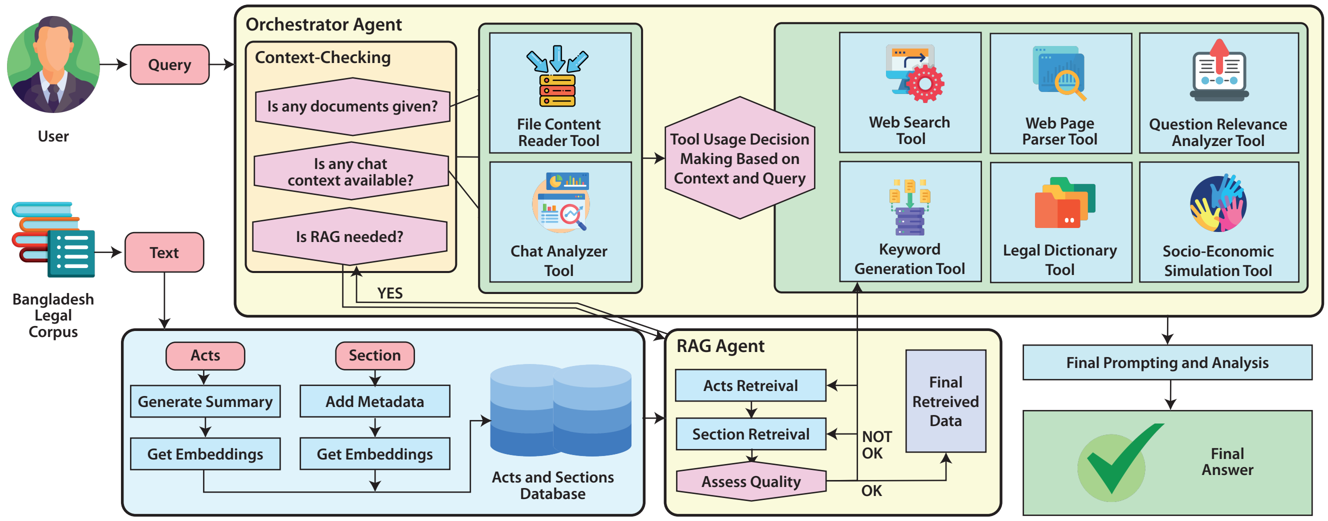

Figure 1: System Architecture and Workflow of our Multilingual Legal Assistant Agent for Bangladesh

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. 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 outlined in Figure 1.

3.1.1. Retrieval-Augmented Generation (RAG)

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

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

■ **Database Development.** Initial experiments showed that naïve retrieval often produced responses by combining content from unrelated Acts. To address this, we developed two separate vector databases. The Act database was constructed by collecting full-text legislative data from the official Bangladesh Law and Justice website. For each Act, we generated an LLM-based summary to create a compact, searchable representation, indexed alongside metadata. The Section database was built by chunking individual sections (1–2 chunks each) and attaching relevant Act metadata for precise filtering during Section-level search.

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

3.1.2. Agents

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

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

■ **RAG Agent** manages the retrieval process end-to-end, from keyword generation to assessing the 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. These are:

1. 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.
2. 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.
3. To facilitate external information access, a DuckDuckGo-based **Web Search** tool retrieves top-ranked results, including titles, URLs, and content snippets.
4. 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.
5. 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.
6. **Chat Analyzer** analyzes previous chat to get context in chat mode.
7. 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.
8. 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 J.

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.

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

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.

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	**<u>35.20</u>**	**<u>70.40</u>**	**<u>78.60</u>**	**<u>81.00</u>**	**<u>36.20</u>**	**<u>71.20</u>**	**<u>79.40</u>**	**<u>81.80</u>**
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	**<u>71.60</u>**	68.00	26.20	60.40	64.20	70.40
Command-A-8B	25.00	**<u>60.20</u>**	71.00	**<u>74.40</u>**	25.20	**<u>73.00</u>**	**<u>74.20</u>**	**<u>76.00</u>**
Llama3.1-8B-Instruct	**<u>28.20</u>**	34.40	46.20	50.00	**<u>30.20</u>**	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	**<u>45.00</u>**	**<u>62.20</u>**	**<u>75.00</u>**	**<u>79.80</u>**	**<u>47.20</u>**	**<u>65.00</u>**	**<u>77.20</u>**	**<u>80.20</u>**

4.2. Evaluation Metrics

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

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

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

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

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**

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 E.

4.3. Baselines and Human Performance

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

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

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

■ **Viva Voce:** The final stage evaluates legal articulation, judgment, and argumentative coherence. Pass rates for this stage are notably high, as the viva is essentially an oral extension of the written exam, and candidates who reach this point are typically well-prepared. In 2022, 96.65% of candidates (5,348 out of

5,533) passed the viva, while in 2023, 97.25% (2,973 out of 3,057; some participants were conditionally allowed) successfully cleared this stage.

5. Results and Findings

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

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

■ **Performance on Written Exam.** Written evaluation results in Table 2 reveal both parallels and divergences from the MCQ setting. Proprietary models dominate, with Gemini-2.5-Flash achieving the highest scores across years, consistently surpassing 80% with tools. Open-source models demonstrate greater competitiveness in written tasks compared to MCQs: large-scale models like Llama3.1-70B and Qwen3-30B reach 79–81% with multi-step RAG and tools. Smaller models, such as Command-A-8B and Qwen3-4B, show significant gains when retrieval is added, highlighting the benefit of structured augmentation. Year-to-year consistency indicates generalization across cohorts rather than memorization. Scaling advantages are more pronounced in written evaluation, where moving to 27B or 70B substantially boosts performance. These results suggest that open-source models, paired with robust RAG pipelines, can provide cost-effective, competitive legal reasoning, especially in Global South contexts like Bangladesh.

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

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

Detailed analysis is available in Appendix H.

6. Error Analysis

■ **MCQ.** As detailed in Appendix F.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 detailed in Appendix F.2, our analysis of model performance on the 2022–2023 Bangladesh Bar Council examination datasets reveals systematic errors stemming from both knowledge gaps and reasoning limitations. The model frequently uses imprecise legal terminology, such as "injury" instead of "damage" under the Specific Relief Act, 1877, omits key statutory conditions, and produces factually incorrect responses, including denying the existence of general exceptions in the Penal Code, 1860. It also conflates civil and criminal law and occasionally hallucinates, generating fabricated cases like *Deowaney Mokdama v. Hazirawala and Garahazira* with spurious procedural details. These errors arise from limited exposure to legal corpora, a fluency-biased generation mechanism, inadequate multi-step reasoning, contextual misinterpretation, and inherent hallucination tendencies in language models. Mitigation strategies include fine-tuning on authoritative legal texts, implementing explicit reasoning frameworks to reduce omissions and domain conflation, and integrating verification against statutory and case law databases to improve factual accuracy. Collectively, these measures aim to enhance the interpretive fidelity and reliability of AI-generated legal responses and highlight the challenges of deploying large language models in high-stakes legal assessment contexts.

7. Evaluator Response Analysis

■ **Written.** As summarized in Appendix G.1, evaluators found that large retrieval-augmented models produced coherent, exam-style answers following the IRAC pattern, with strong statutory recall and clear drafting. Their responses often mirrored human exam scripts, combining issue identification, legal statement, and concise application. However, omissions of procedural prerequisites, such as conditions for injunctions or receiverships, were frequent and penalized as substantive errors. Arithmetic failures in limitation-period calculations and incomplete synthesis across the CPC, Evidence Act, and Specific Relief Act were also noted. Evaluators highlighted bilingual clarity and structured reasoning as major strengths but criticized inconsistent statutory precision and occasional hallucinated case citations. Overall, the written analysis shows exam-ready compositional fluency yet emphasizes the need for deterministic procedural calculators, statutory-fusion modules, and fine-tuning on annotated bilingual exam data to achieve distinction-level precision.

■ **Viva.** As detailed in Appendix G.2, viva evaluations revealed that Mina performed with composure, clear diction, and accurate statutory articulation under conversational pressure. Examiners praised its coherence and ability to paraphrase complex doctrines into accessible explanations, though it sometimes failed to adapt reasoning when factual variations were introduced. A key weakness was the lack of clarifying questions and the flattening of hierarchical reasoning, leading to incorrect procedural advice in dynamic exchanges. Rapid doctrinal checks exposed occasional section misattributions, which evaluators treated as major oral faults. Feedback emphasized that distinction-level performance requires interpretive flexibility, acknowledgment of uncertainty, and deeper policy reasoning beyond rule recital. System-level remedies include integrating a clarification policy, dialogue-state tracking, and real-time procedural validation to enhance adaptive accuracy and examiner trust during live oral assessment.

8. Cost-per-Query Comparison and Affordability Implications

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

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

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

9. Concluding Remarks

In this paper, we present Mina, a multilingual LLM-based legal assistant tailored to the Bangladeshi legal context, designed to improve access to justice for low-income and linguistically diverse populations. By integrating two-stage RAG, multilingual embeddings, and selective tool augmentation, the system delivers legally coherent, context-aware responses in both Bengali and English across multiple exam modalities, including MCQs, written, and oral viva assessments. Even mid-sized open-source models approach or exceed average human performance, while top-tier models like Gemini-2.5 Flash achieve near-ceiling results.

The system's modular and lightweight architecture enables efficient deployment in resource-constrained environments, while its bilingual and dictionary-enhanced capabilities allow accurate interpretation of colonial-era and Farsi-influenced terminology. Detailed error analysis reveals challenges in hierarchical procedural reasoning, timeline computation, and adaptive oral exchanges, highlighting opportunities for structured procedural modules, real-time verification, and dialogue-state tracking. Our findings suggest that retrieval alone is insufficient for high-stakes legal tasks; targeted fine-tuning, contrastive training, and interactive clarification policies can further improve reliability. Overall, Mina demonstrates that careful system design, rigorous evaluation, and domain-specific adaptation enable LLMs to provide practical, scalable, and cost-effective legal assistance. Beyond Bangladesh, this approach offers a template for low-resource, multilingual legal environments worldwide, illustrating the potential of AI to enhance equitable access to legal knowledge and support professional legal workflows while maintaining human oversight.

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.

Reproducibility Statement

To ensure full reproducibility of our experiments, all code, trained retrieval indices, and processed datasets will be made publicly available upon acceptance. Detailed instructions for setting up the environment, running the Orchestrator Agent, executing the two-stage RAG pipeline, and reproducing all evaluations for MCQs, written exams, and viva voce are included in the paper and repository. Additionally, prompts, scoring rubrics, and evaluation protocols are documented to allow independent verification of both quantitative results and qualitative analyses.

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.

References

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

Farzana Akter. 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.

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

Tanim Asjad. 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.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan

Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025. URL https://arxiv.org/abs/2504.00698.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 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.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 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.

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

Jens Frankenreiter and Julian Nyarko. *Natural Language Processing in Legal Tech*, page 70–90. Cambridge University Press, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor

Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad

Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Péter Homoki and Zsolt Ződi. 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.

Jia-Hong Huang, Chao-Chun Yang, Yixian Shen, Alessio M. Pacces, and Evangelos Kanoulas. Optimizing numerical estimation and operational efficiency in the legal domain through large language models, 2024. URL https://arxiv.org/abs/2407.19041.

Atin Sakkeer Hussain and Anu Thomas. Large language models for judicial entity extraction: A comparative study, 2024. URL https://arxiv.org/abs/2407.05786.

Md Jahirul Islam, Masahiro Suzuki, and Nurunnahar Mazumder. 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.

Arafat Reza Jaan. 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.

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

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. Leveraging large language models for learning complex legal concepts through storytelling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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.

Md Ibrahim Khalilullah. 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.

Shammem Ara Knock, Dr. Md. Omar Faruk, Rukhsana Siddiqua, and Asma Akter. 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.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. Large language models in law: A survey, 2023. URL https://arxiv.org/abs/2312.03718.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. Legalagentbench: Evaluating llm agents in legal domain, 2024. URL https://arxiv.org/abs/2412.17259.

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

Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. 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.

Marios Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, and Konstantinos Tarabanis. A large language model based legal assistant for governance applications, May 2024a. URL osf.io/94fmq.

Marios Evangelos Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, and Konstantinos Tarabanis. *A Large Language Model Agent Based Legal Assistant for Governance Applications*, page 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.

Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. Better call gpt, comparing large language models against lawyers, 2024. URL https://arxiv.org/abs/2401.16212.

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

Arpeeta Shams Mizan. 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, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen

Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, Shivalika Singh, Seraphina Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh Fadaee, and Sara Hooker. The multilingual divide and its impact on global ai safety, 2025. URL https://arxiv.org/abs/2505.21344.

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

Md. Alamgir Sarkar Raj. 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.

Laura Safdie. 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.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. Lawluo: A chinese law firm co-run by llm agents, 2024. URL https://arxiv.org/abs/2407.16252.

Ummey Sharaban Tahura and Shawkat Alam. 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-bangladesh-3874031. Accessed: 2025-06-23.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le

Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi

Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar,

Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025a. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat

Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025b. URL https://arxiv.org/abs/2503.19786.

Azmine Toushik Wasi, Wahid Faisal, Mst Rafia Islam, and Mahathir Mohammad Bappy. Exploring possibilities of ai-powered legal assistance in bangladesh through large language modeling, 2024. URL https://arxiv.org/abs/2410.17210.

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

Qiong Yan. 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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Sabina Yesmin. 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.

Zheng-Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen H. Bach, and Julia Kreutzer. The state of multilingual llm safety research: From measuring the language gap to mitigating it, 2025. URL https://arxiv.org/abs/2505.24119.

Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuanjing Huang, and Zhongyu Wei. Multi-agent simulator drives language models for legal intensive interaction. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages

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/.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. How does NLP benefit legal system: A summary of legal artificial intelligence. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 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 and Demartini, 2025). Tools like ROSS Intelligence and Westlaw Edge automate research and e-discovery, while AI chatbots like DoNotPay assist with document drafting. Recent developments integrate large language models (LLMs) with logic programming and multi-agent frameworks leveraging knowledge graphs and Retrieval-Augmented Generation (RAG) to improve reliability and contextual accuracy. Furthermore, LLM-based methods employing specialized prompts and curated datasets for LegalAI tasks (Huang et al., 2024), as well as models like Mistral and Gemma for judicial entity extraction (Hussain and Thomas, 2024), demonstrate superior efficiency and cost-effectiveness over traditional practices (Sun et al., 2024). RAG and agent-based approaches enhance governance and long-form legal question answering (Mamalis et al., 2024a), while storytelling and prompt templates aid complex concept processing and text structuring (Jiang et al., 2024, de Kinderen and Winter, 2024). However, LLM limitations, such as stochastic outputs, hallucinations (Dahl et al., 2024), and fact-checking gaps (Mik, 2024) necessitate more interpretable, evidence-based models (Louis et al., 2024), especially for domain-sensitive contexts like law.

A.2. Agent-focused Legal NLP Research

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

A.3. Multilingual Legal NLP Research

Multilingual NLP research, crucial for low-resource languages, faces challenges from data scarcity, script variations, and cultural nuances (Yong et al., 2025, Peppin et al., 2025). Despite multilingual models like GPT-4 showing promise, performance degradation persists for underrepresented languages such as Bengali. Initiatives like MultiLegalPile (Qin and Sun, 2024) and legal text comprehension studies (Martin et al., 2024, Wehnert, 2023, Homoki and 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 and 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. 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.

C. System Demonstration

Figure 3 shows a demonstration of our system and UI.

D. 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.

D.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.

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)

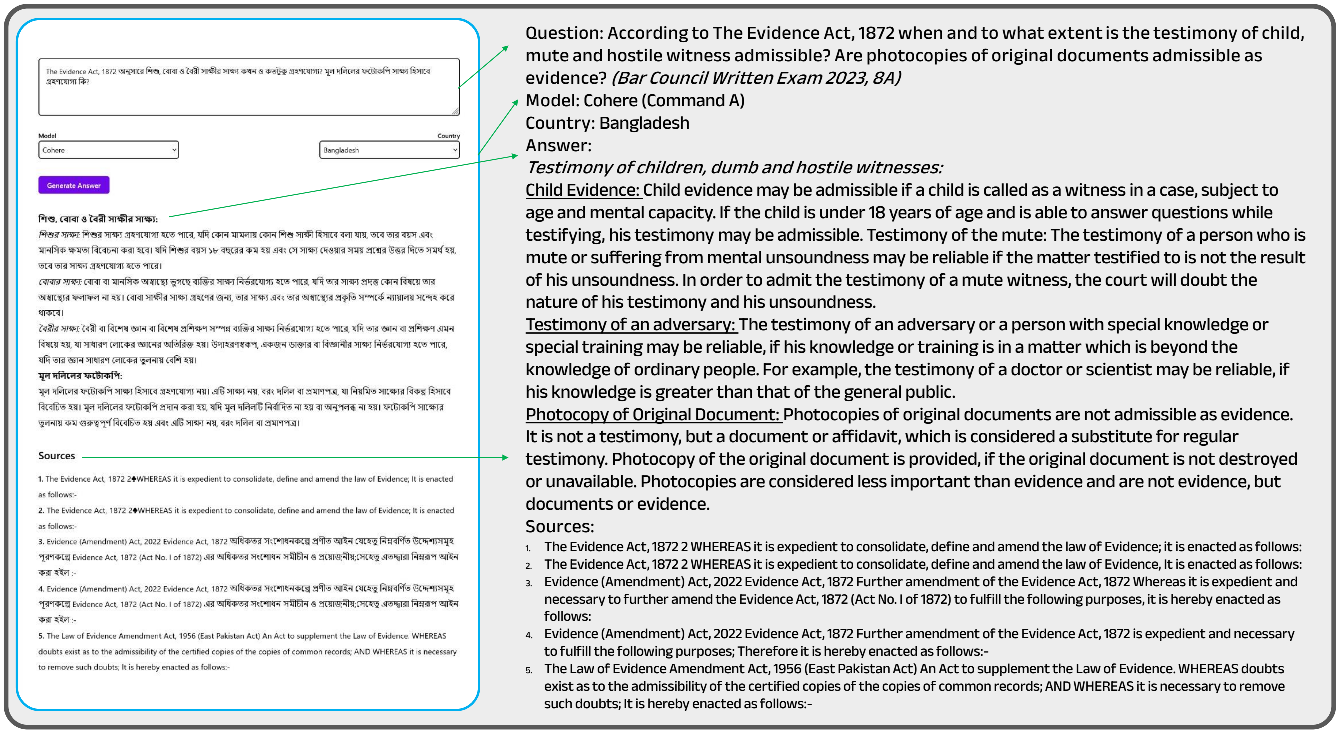

Figure 3: System Demonstration: UI and deployable system of MINA.

D.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.

D.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.

D.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.

D.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.

D.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.

D.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.

E. Additional Information on Experiments and Evaluation

E.1. Evaluators

The evaluation of written and viva responses was conducted by law faculty members from leading Bangladeshi universities. For the viva voce, five evaluators independently scored each response, while three evaluators assessed the written examination. The evaluators applied the same criteria employed in real Bar Council examinations, including accuracy, clarity, contextual understanding, and legal reasoning. No additional training was provided, as the evaluators were already familiar with the official standards and marking rubrics. This approach ensured that the assessment closely mirrors real-world examination practices while maintaining consistency, fairness, and reliability across both the written and oral components.

E.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 G.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.

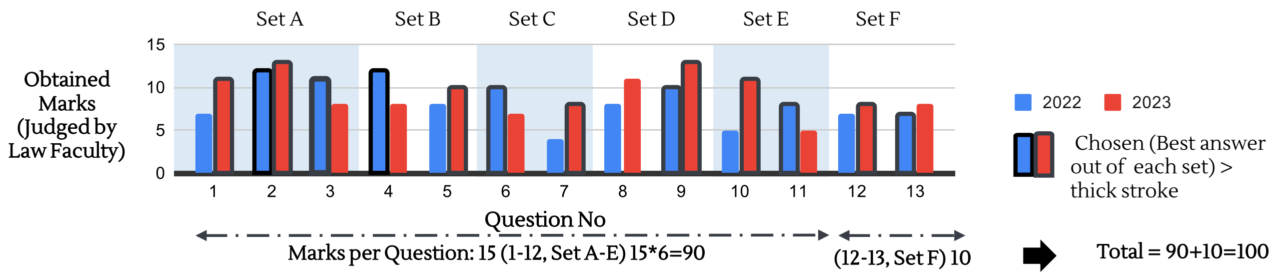

Figure 4: Breaking Down a Written Full Answer (Command-A, Two Step; Examiner 2)

E.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.

E.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 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. Detailed Error Analysis

F.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.

F.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.

F.2.1. Error Types

Here we categorize and discuss the error types:

- **Inaccurate Legal Terminology:** The model frequently substitutes imprecise or informal terms for

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 দায়ের 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)

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

F.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.

F.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.

G. Detailed Evaluator Response Analysis

G.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.

G.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. Extended Analysis

H.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% → 25.2% in 2022 (+17 pts) and 11.2% → 23.4% in 2023 (+12.2 pts). Gemini-2.5-Flash improved from 30.2% → 68.8% (+38.6 pts) in 2022 and 32.4% → 69.2% (+36.8 pts) in 2023, indicating that strong models already leverage latent context. Some regression in 2023 (e.g., Gemini-2.0-Flash: 61.2% → 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 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% → 47.0% in 2022 and 23.4% → 49.2% in 2023. Gemma-3-12B-it improves 35.2% → 48.4% (2022) and 36.2% → 52.4% (2023). Even top-tier Qwen3-30B-A3B-Instruct-2507 increases from 50.4% → 65.6% (2022) and 52.4% → 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% → 70.8% in 2022 and 67.2% → 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% → 49.2%), while top models plateau (Gemini-2.5-Flash 75.6% → 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.

H.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.

H.3. Detailed Findings from Viva Voce (Oral)

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

■ **Progressive Setup Sophistication Reflects Realism and Context Awareness** Two-Step RAG, which filters and reranks relevant materials, substantially boosts performance for mid-tier and large models: Command A moves 55.6 → 70.4, Gemma-3-4B-it 50.4 → 65.2, and Gemini-2.5-Flash 70.4 → 79.2. Gains derive from the model's improved ability to prioritize context most pertinent to the question, reduce hallucinations, and apply statutes in sequence during oral reasoning. Cross-model patterns show mid-tier models benefit disproportionately, while top-tier models plateau (Llama3.1-70B-Instruct 63.2 → 77.2). This trend underscores that setup sophistication amplifies latent reasoning but does not create capability de novo. Noise reduction and contextual relevance emerge as key drivers for performance in real-world oral exams, where follow-up probing and adaptive reasoning are required.

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

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

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

I. 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.

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

J. Prompts

J.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.

---
```

J.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}

---
```

J.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)

---
```