AI for Legal Domain Identification and Guidance in Sri Lankan Law

Anushka Athulathmudali[1]

Publication Date: 2025/12/15

Abstract: Sri Lanka's pluralistic legal system is difficult for many citizens to navigate, leading to missed deadlines and reduced access to justice. This research investigates whether Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) can provide accurate, cost-efficient civil-law guidance tailored to Sri Lanka. A curated dataset of expert-validated legal scenarios was developed and integrated into a modular RAG pipeline. Two LLM backends, GPT-3.5-Turbo and Mistral-7B-v0.1 were evaluated under identical conditions for legal accuracy, latency, and cost. Results show that GPT-3.5-Turbo achieved the best overall performance with 92.5% accuracy, 4.17s latency, lowest cost per correct response, making it suitable for responsive public-facing legal-information services. Mistral-7B-v0.1 demonstrated competitive accuracy of 82.5% with full data-sovereignty benefits, but higher latency limits interactive deployment, better aligning it with institutional environments prioritizing privacy and local infrastructure control. The study provides a practical framework for responsible legal-AI adoption in Sri Lanka and contributes a reusable RAG architecture and localized dataset. These findings suggest that AI-enabled legal guidance can support access-to-justice goals when deployed with appropriate safeguards and governance.

Keywords: *Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Legal Corpus, Procedural Accuracy Evaluation, Latency and Cost Profiling, GPT-3.5-Turbo; Mistral-7B, Legal Hallucination Mitigation, Access to Justice in Sri Lanka.*

How to Cite: Anushka Athulathmudali (2025) AI for Legal Domain Identification and Guidance in Sri Lankan Law. *International Journal of Innovative Science and Research Technology*, 10(12), 609-621.
https://doi.org/10.38124/ijisrt/25dec441

I. INTRODUCTION

The legal system of Sri Lanka has a pluralistic foundation, integrating elements from Roman-Dutch law, English common law, parliamentary statutes, and indigenous customary laws. It is rich but creates a complex environment that many citizens struggle to navigate [1]. Individuals encountering civil issues often find it difficult to identify the relevant legal domain or understand the procedural steps required, which can result in missed deadlines and lost legal remedies. For example, a fundamental rights petition must be filed at the Supreme Court within one month [2], and failure to meet this requirement can prevent a case from being heard. These obstacles result in severely constrained access to justice, as evidenced by survey data showing only 19% of Sri Lankans with a legal problem obtained help, while 8% abandoned their efforts entirely [3].

Advances in Large Language Models (LLMs) [4] and Retrieval-Augmented Generation (RAG) techniques [5] offer a potential solution by enabling AI systems to generate grounded, context-aware guidance. However, applying these technologies to the Sri Lankan legal domain presents two core challenges. First, general-purpose LLMs lack jurisdiction-specific knowledge and are not trained on Sri Lankan legal procedures [6]. Second, the choice between proprietary and open-source models involves significant trade-offs. Proprietary models such as GPT offer strong performance but require external data processing and incur ongoing costs [7][8], while open-source models such as Llama and Mistral enable local deployment with better data control but may not achieve comparable accuracy for specialized legal tasks.

The resulting research problem is to determine which type of RAG-LLM configuration, proprietary or open-source, offers the most reliable, cost-effective, and practical foundation for civil-law guidance in Sri Lanka. Addressing this gap is essential for building AI systems that are both trustworthy and operationally feasible within a low-resource, high-stakes legal environment.

This research aims to design and evaluate an AI-based legal guidance system capable of performing two key functions: 1) classifying user queries into appropriate legal domains, and 2) generating procedurally accurate guidance.

To support these tasks, the study constructs a curated corpus of Sri Lankan civil-law scenarios and builds a RAG pipeline that retrieves relevant legal content and generates grounded responses. Multiple LLM backends including leading proprietary models and open-source alternatives were integrated into this pipeline to enable a controlled comparison.

The contributions of this study extend to both practice and research. Practically, it demonstrates how AI systems can provide accurate, low-cost preliminary legal guidance, potentially improving public understanding of legal processes and reducing barriers to justice. Academically, it offers a comparative evaluation of LLM performance in a low-resource legal domain and introduces a structured dataset specific to Sri Lankan civil law. The study also examines operational considerations such as latency, cost, privacy, and data sovereignty, which are critical for sustainable deployment.

The proposed solution involves implementing a RAG-based system across several LLM configurations and evaluating them using metrics focused on classification accuracy, procedural correctness, latency, and operational cost. The resulting empirical analysis identifies the trade-offs inherent in each configuration, offering evidence-based insight into whether open-source models can reliably support legal guidance or whether proprietary systems remain essential for this domain.

II. LITERATURE REVIEW

➢ *Legal AI Systems*

AI-powered legal chatbots and virtual assistants have emerged across jurisdictions to improve access to legal information, support document drafting, and handle basic legal queries.

- *Types and Examples of Legal Chatbot Systems*

Legal chatbots generally fall into two main categories [9]. The first consists of rule-based or decision-tree systems, which rely on predefined rules to guide users through structured interaction pathways. These systems are cost-effective and work well for predictable, task-specific queries, but they cannot process free-text descriptions or manage complex, ambiguous issues, limiting their usefulness in dynamic real-world scenarios [10]. The second category comprises machine-learning-based chatbots, which use statistical or generative models to interpret natural-language input and provide more flexible, context-aware responses [9]. These systems can better handle variability in user queries and address tasks that extend beyond the rigid structure of rule-based approaches [10].

These models support natural-language queries and generate context-aware responses. Examples include DoNotPay, which assists with consumer law issues and small-claims disputes [11], and Ailira, which supports tax, estate, and business law and can generate legal documents [12]. More advanced systems include ChatLaw, an open-source legal LLM that uses hybrid retrieval for improved accuracy [13]. Its enhanced version employs a Mixture-of-Experts architecture, multi-agent design, and knowledge graphs, outperforming GPT-4 on LawBench [14]. Other systems include LawBot (Elexirr), which predicts outcomes of criminal cases [15], and proprietary tools such as Lexis+ AI and Westlaw's AI-Assisted Research, which integrate internal RAG pipelines to improve research and summarization [16].

- *Core Limitations and Risks*

General-purpose LLMs demonstrate high hallucination rates on legal tasks, ranging from 58% to 82% [17]. Even specialized legal RAG systems produce hallucinations in 17% to 33% of cases [16], indicating persistent reliability challenges.

Jurisdictional adaptation remains a major obstacle. The failure of the JuridiQC chatbot in Quebec illustrates the difficulty of adapting AI to local legal frameworks and maintaining the boundary between "legal information" and "legal advice" [18]. LLMs also struggle with maintaining conversational context [19] and suffer from the black-box problem, making their reasoning opaque and raising professional liability concerns [20].

Bias amplification presents additional risks, as LLMs inherit societal biases embedded in their training data and can inadvertently reinforce stereotypes [21]. Moreover, many legal chatbots are restricted to narrow subdomains, limiting their usefulness for complex queries [12]. Finally, privacy and data governance remain significant concerns, as legal chatbots often process sensitive personal information requiring strict compliance measures [22].

➢ *RAG Architectures*

Retrieval-Augmented Generation (RAG) enhances language models by conditioning outputs on external documents rather than relying solely on parametric knowledge. This reduces hallucinations and improves factual grounding across knowledge-intensive tasks [5].

A standard RAG pipeline includes: a document index (e.g., FAISS), a retriever (e.g., BM25 or DPR), and a generator that synthesizes responses using retrieved passages. Dense Passage Retrieval (DPR) is widely used and consistently outperforms sparse retrieval in semantic tasks [23].

Architectures such as Fusion-in-Decoder (FiD) achieve strong performance by encoding retrieved passages independently and integrating them only at decoding time, allowing efficient evidence aggregation at scale [24]. Retrieval-aware approaches like REALM further demonstrate improvements in factual accuracy by combining pretrained language models with learned retrievers [24]. In legal contexts, Retrieval-Augmented Generation (RAG) offers two major advantages: grounding, which provides traceable citations through explicit retrieval, and updatability, as the underlying corpus can be refreshed without retraining the language model. However, retrieval errors and weak passage alignment can still lead to misleading or incorrect outputs, meaning that RAG mitigates, but does not eliminate hallucinations [5].

Legal-specific retrieval evaluation focuses on fine-grained snippet extraction and citation-traceable relevance, as demonstrated by the LegalBench-RAG benchmark. In retrieval design, key considerations include effective chunking strategies, retriever selection, and evaluation using precision and recall across multiple datasets [25]. Dense retrieval methods such as DPR have shown consistent performance gains over sparse baselines in semantic matching tasks [23].

➢ *LLM Evaluation Methods: How to Measure Legal Accuracy*

Evaluating legal accuracy requires combining retrieval metrics, generation faithfulness, citation correctness, and expert judgment. No single metric captures all aspects, so legal evaluations typically rely on multi-dimensional frameworks [26].

Retrieval evaluation uses metrics such as Recall@K, Precision@K, and MRR. Legal benchmarks emphasize minimal-span retrieval due to the precision required in statutory and case-law interpretation [27].

Generation accuracy and faithfulness are measured using factuality and claim-level metrics, including entailment checks, fact verification, citation matching, and provenance-aware scoring such as KILT [26]. Hallucination detection methods help flag risky outputs but cannot replace expert review [28].

Citation and provenance accuracy are essential for legal contexts. Evaluations check whether generated citations are correct, contextually appropriate, and directly linked to retrieved text [27].

Best practice involves combining automatic metrics with adversarial tests and human expert assessment. Benchmarking libraries such as BERGEN support full-pipeline evaluation across RAG configurations [29]. Evaluations must also document dataset scope, statutory versioning, and assessor qualifications, as operational considerations like latency, cost, and provenance trails significantly impact practical deployment [25].

➢ *Open-Source vs Proprietary LLMs: Performance Trade-offs in Specialized Domains*

Selecting between open-source and proprietary LLMs involves weighing accuracy, cost, privacy, customization, and infrastructure demands. Studies such as ContractEval show that proprietary models outperform open-source ones on legal risk identification, though fine-tuned open models can be competitive in narrow tasks [7]. LawBench similarly finds GPT-4 superior in legal reasoning, while fine-tuned open models close the gap in classification and extraction tasks [30].

Cross-domain research shows that fine-tuned open-source models can outperform larger proprietary systems in highly specialized tasks, such as molecule editing in TOMG-Bench [31]. Machine reading comprehension evaluations confirm proprietary models' dominance but highlight the cost and latency benefits of open-source alternatives in resource-constrained environments [32].

Proprietary models offer strong general reasoning and consistent latency but rely on cloud infrastructure and restrict fine-tuning. Open-source models offer full data control, local deployment, and customization, but require hardware investment and ongoing maintenance.

- *Accuracy and Task Complexity*

  Proprietary models excel in complex reasoning tasks across law and biomedicine [30][31], while open-source models perform competitively on narrow or structured tasks, particularly with domain-specific fine-tuning [31][32].

- *Model Size and Scaling*

  Performance gains plateau beyond mid-sized models; domain tuning and high-quality data often matter more than parameter count [7][30][32].

- *Fine-Tuning and Specialization*

  Fine-tuning is essential for open-source performance, yet even tuned models may lag behind highly capable proprietary models. Instruction tuning helps close the gap [7][30][32].

- *Cost and Resource Requirements*

  Open-source models provide long-term cost efficiency but require significant upfront infrastructure, unlike pay-per-use proprietary APIs [7][32].

- *Privacy and Legal Risk*

  Local deployment of open-source models supports confidentiality and jurisdiction-specific compliance, particularly important in sensitive domains like law and biomedicine [7][30][31].

- *Latency and Maintenance*

  Proprietary models deliver optimized latency; open-source deployment requires engineering trade-offs and resource management. Quantization reduces computational costs but can reduce accuracy, especially on reasoning tasks [7][30][32].

- *Generalization vs Domain Accuracy*

  General-purpose LLMs excel in broad reasoning, while fine-tuned domain-specific models often surpass them in structured extraction tasks [31][32]. The success of legal benchmarks like LawBench underscores the importance of jurisdiction-specific tuning [30].

➢ *Sri Lankan Legal Informatics*

Sri Lanka's digital legal information ecosystem includes official, commercial, and public-access resources, but coverage and structure remain fragmented. LawNet provides consolidated legislation and selected case law [33], while the Court of Appeal publishes recent judgments [34]. Government portals host Gazettes, Bills, and statutory instruments, often as non-machine-readable PDFs [35].

Supplementary resources include the Laws of Sri Lanka consolidated statutes [36], CommonLII's collections of statutes and judgments [37], and commercial platforms such as vLex offering curated legal databases [38]. International rule-of-law assessments highlight demand for better legal information access [39].

Despite this variety, significant obstacles persist: paywalls, licensing restrictions, inconsistent formatting, and limited machine-readable data. Scholarly analysis confirms that Sri Lanka lacks sufficient structured legal data for training robust AI systems [6].

There is, however, growing momentum toward digitization. A recent proposal outlines an AI-enabled framework for the judiciary, including centralized digital records, OCR-based document processing, and AI-driven workflow automation to address case backlogs [40]. This signals a move toward a more integrated and AI-ready legal data ecosystem.

➢ *Research Gap Identification*

Global advances in legal AI such as DoNotPay, Ailira, and ChatLaw and extensive literature on RAG architectures and LLM evaluation demonstrate the potential of AI-driven legal guidance systems. Benchmarks like LawBench and ContractEval show that proprietary models excel in reasoning, while open-source models offer advantages in cost, privacy, and customizability.

However, these developments have not been adapted to Sri Lanka's unique legal context. Existing legal AI systems are tailored to jurisdictions with structured, machine-readable datasets, unlike Sri Lanka's mixed system of Roman-Dutch law, English common law, local statutes, and customary laws. Domestic legal resources, though numerous, are fragmented and often unsuitable for direct computational use.

No existing research has constructed a curated Sri Lankan civil law corpus, built a RAG pipeline upon it, and comparatively evaluated proprietary and open-source LLMs for legal accuracy, latency, and cost. This study addresses this gap by determining which RAG configuration provides the most practical and reliable solution for Sri Lanka's civil-law context, offering an evidence-based blueprint for responsible legal AI development.

III. RESEARCH DESIGN AND METHODOLOGY

➢ *Dataset Construction*

Given the limited availability of structured and machine-readable Sri Lankan legal data, this study adopts an expert-curated dataset approach. Datasets built solely from raw legislation or case law often contain OCR noise, missing metadata, and lack procedural context, which can lead to inaccurate retrieval and hallucinated guidance [41]. Expert-crafted records instead reflect authentic citizen language and ensure legal accuracy and completeness.

Similar practices are standard in legal NLP. Benchmarks such as LegalBench [25][42], ACORD [43], and ContractEval [7] demonstrate that domain expertise improves annotation fidelity and the legal reliability of downstream evaluations. Following this precedent, a curated dataset of Sri Lankan civil-law scenarios was developed.

- *Dataset Design*

A small panel of Sri Lankan legal practitioners designed and validated matter records across priority civil-law domains, including family law, fundamental rights, and succession. This approach aligns with established legal NLP benchmarks, where domain expertise is essential to ensure annotation fidelity and legal reliability [7][25][42][43]. The resulting corpus contains 72 records, each describing a realistic scenario (100–300 words) in everyday language. Each record includes:

✓ A clear domain label
✓ Procedural context such as jurisdiction, filing requirements, and timelines
✓ Guidance on relevant evidence and potential remedies

This structure supports both effective retrieval and accurate evaluation of legal reasoning.

- *Integration within System Architecture*

Each record is stored as a lightweight Document object (JSON/NoSQL format) with minimal metadata. Documents are converted to vector embeddings and indexed in the RAG retrieval system to serve as the authoritative knowledge base. A held-out test set is used for evaluation, and legal experts perform binary accuracy scoring to confirm procedural correctness.

- *Sample Size Considerations and Ethical Compliance*

Although the dataset size limits large-scale model fine-tuning, it is adequate for controlled evaluation of RAG configurations. To account for limited sample size, results combine quantitative performance metrics with qualitative error analysis.

All data is hypothetical or anonymized and reviewed under informed consent. Outputs are explicitly positioned as legal information, not legal advice, in compliance with ethical standards and Sri Lanka's Personal Data Protection Act of 2022.

➢ *System Architecture: RAG Pipeline Design*

The system uses a Retrieval-Augmented Generation (RAG) architecture, combining semantic retrieval with LLM-based response generation to improve factual grounding and mitigate hallucinations [5]. By linking the model's responses to credible and context-specific data stored in a structured legal knowledge base, this architecture effectively reduces a common issue in standalone LLMs producing inaccurate or unverified information.

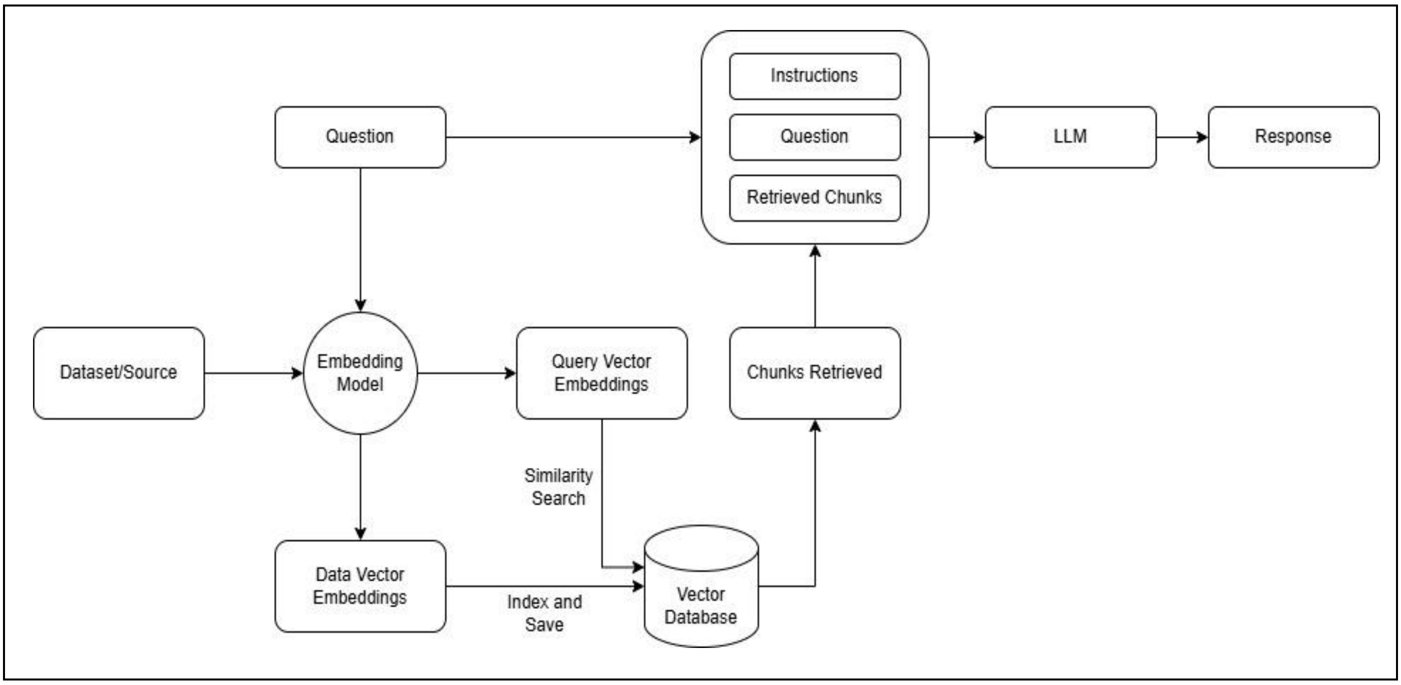

Fig 1 System Architecture of the RAG Pipeline for the Proposed AI Legal Assistant.

Fig 1 presents the end-to-end data workflow, beginning with document ingestion and vector embedding, followed by retrieval and augmentation, and concluding with the generation of the final response. As shown, the RAG pipeline is designed as a step-by-step modular system with four main phases: ingestion, retrieval, augmentation, and generation, ensuring a systematic flow of information from the knowledge base to the LLM-powered output.

Knowledge Ingestion: The expert-curated legal scenarios are encoded into semantic vectors using the text-embedding-3-small model and stored in a Chroma vector database, which supports scalable similarity search and metadata filtering [44].

Retrieval: User queries are embedded into the same vector space and compared against the index. The system retrieves the top k relevant records (k = 3), enabling interpretation of queries expressed in everyday, non-technical language and supporting dense semantic retrieval performance advantages over sparse methods [23].

Grounded Generation: Retrieved content is injected into a structured prompt that constrains the model to rely solely on the provided legal context, enhancing traceability and discouraging unsupported assumptions [5][45]. GPT-3.5-Turbo is used as the generator with temperature set to 0 to ensure deterministic outputs suitable for legal information tasks.

This modular architecture allows future expansion to additional legal domains, integration of more advanced retrieval models, and adaptation to varied deployment environments.

➢ *Model Selection and Rationale*

A dual-model evaluation was adopted to compare proprietary and open-source LLM configurations under identical retrieval and prompting conditions, isolating the language model as the primary performance variable.

Mistral-7B v0.1 was selected as the open-source model due to its strong performance relative to similarly sized models, outperforming Llama-2-13B on its release benchmarks [46]. Architectural innovations including grouped-query attention (GQA) and sliding-window attention (SWA) improve context handling and inference efficiency [46], which are critical properties for RAG systems. Its permissive Apache 2.0 license additionally supports academic experimentation and deployment without proprietary constraints.

GPT-3.5-Turbo was chosen as the proprietary model based on its robust instruction-following ability and strong applied performance in machine-reading tasks, where it achieves state-of-the-art accuracy and semantic comprehension [32]. The model also offers optimized latency suitable for interactive systems, though reliance on a paid API introduces cost and cloud-based data handling considerations.

This controlled comparison enables a focused analysis of differences in legal accuracy, operational efficiency, and cost, attributable specifically to model architecture and deployment modality.

➢ *Evaluation Framework*

A three-metric evaluation framework is used to assess real-world suitability,

- Accuracy and Procedural Correctness: Performance is assessed through legal expert review of domain classification and procedural validity. Scoring evaluates the inclusion and correct sequencing of required legal steps and the avoidance of hallucinations, consistent with

legal-faithfulness and expert-grounded evaluation practices in current literature [25][26][28][42].
- Latency: Measured as end-to-end response time, capturing the combined impact of retrieval and generation. Latency reflects practical usability under realistic user interaction conditions and is a key operational metric in legal and machine-reading system deployment [7][32].
- Cost Efficiency: GPT-3.5-Turbo is evaluated using official per-token API pricing, while Mistral-7B-v0.1 cost reflects GPU compute time associated with open-source deployment. This highlights trade-offs between subscription-based proprietary services and infrastructure-dependent self-hosting of open-source models [7].

This framework enables a holistic comparison of legal accuracy, interactive responsiveness, and resource requirements key determinants of deployability in low-resource legal environments.

➢ *Experimental Setup*

A 20% held-out test set was reserved exclusively for evaluation to prevent data leakage. Both RAG pipelines share the same embedded corpus, Chroma vector index, and structured prompt format, ensuring a valid controlled comparison where the LLM is the sole experimental variable.

The GPT-3.5-Turbo configuration uses the OpenAI API with temperature set to 0 for deterministic, reproducible output. Standard API constraints apply, including external cloud processing and token-based billing.

The Mistral-7B-v0.1 configuration is executed locally on Google Colab Pro with GPU acceleration via HuggingFace Transformers. This reflects a practical, replicable environment that balances performance requirements and resource accessibility for open-source deployment.

All model responses are logged automatically and evaluated against accuracy, latency, and cost efficiency, followed by expert-based qualitative error analysis to assess legal reliability.

➢ *System Implementation*

The system was implemented using a modular Retrieval-Augmented Generation (RAG) pipeline integrated with interchangeable LLM backends. The curated legal scenarios were encoded using the text-embedding-3-small model and stored in a Chroma vector database to support fast vector similarity search and provenance-aware retrieval. Each scenario represents a self-contained civil-law case summary, enabling precise retrieval at the level of actionable legal guidance.

For each user query, the system retrieves the top-3 semantically similar records based on cosine similarity and injects them into a structured legal-assistant prompt that enforces grounding, procedural clarity, and cautious language. GPT-3.5-Turbo and Mistral-7B-v0.1 are evaluated under identical retrieval and prompting conditions, isolating the language model as the primary experimental variable. All experiments were executed in Google Colab Pro with GPU acceleration for the open-source model and API-based inference for GPT-3.5-Turbo. Automatic logging captured retrieved sources, generated responses, latency, and token or compute usage to support evaluation of accuracy, responsiveness, and cost.

This design ensures a fair and controlled comparison of system behavior across deployment models while maintaining auditability and replicability in a resource-constrained environment.

IV. RESULTS AND ANALYSIS

The performance of GPT-3.5-Turbo and Mistral-7B-v0.1 was evaluated using a 20% held-out test set containing 40 representative civil-law queries. Evaluation focused on three deployment-critical dimensions in legal-AI systems: 1) accuracy, 2) latency, and 3) economic cost.

This multi-metric framework reflects best practice guidance that correctness, responsiveness, and affordability jointly determine real-world viability [47].

GPT-3.5-Turbo demonstrated the strongest accuracy, answering 37 out of 40 scenarios correctly (92.5%), while Mistral-7B-v0.1 achieved 82.5% accuracy. Errors and empty responses were counted equally to capture operational reliability. The results are summarized in Table 1. Although accuracy differences were not statistically significant at the 95% confidence level, the 10-percentage-point gap is nevertheless practically relevant for public-facing legal information services where the cost of an incorrect answer may be high.

Table 1 Accuracy Performance

Model	Correct / Total	Accuracy (%)
GPT-3.5-Turbo	37 / 40	**92.5**
Mistral-7B-v0.1 (L4 GPU)	33 / 40	82.5

Latency exhibited the most pronounced performance disparity. GPT-3.5-Turbo, hosted in a highly optimized cloud environment, achieved an average response time of 4.17 seconds per query well within the threshold required for interactive conversational systems. Mistral-7B-v0.1 performed substantially slower, with latency dependent on GPU configuration. The L4 GPU was ultimately selected for comparison, producing an average response time of 15.64 seconds. Although performance on the A100 GPU was faster

(9.25 seconds), it required significantly greater compute resources. Table 4.2 summarizes the observed values.

Table 2 Average Latency per Query

Model	Infrastructure	Avg. Latency (s)
GPT-3.5-Turbo	API (Cloud)	**4.174**
Mistral-7B-v0.1	L4 GPU	15.644
Mistral-7B-v0.1	A100 GPU	9.252
Mistral-7B-v0.1	T4 GPU	85.296

Thus, GPT-3.5-Turbo was approximately 73% faster than Mistral-7B-v0.1 under its selected deployment setting. In practice, this difference corresponds to much smoother user experience in systems requiring quasi-real-time conversation.

Cost results further reinforce the advantage of GPT-3.5-Turbo. Based on OpenAI's token pricing [48] and Colab Compute Unit billing [49], the cloud-hosted model achieved the lowest per-query cost at USD 0.000487. Mistral-7B-v0.1 cost varied with hardware and on the selected L4 GPU averaged USD 0.000742, reflecting the higher latency and compute requirements. Table 4.3 summarizes the full cost comparison.

Table 3 Cost Per Query

Model	Infrastructure	Avg. Cost (USD)
GPT-3.5-Turbo	API	0.000487
Mistral-7B-v0.1	L4 GPU	0.000742
Mistral-7B-v0.1	A100 GPU	0.000439
Mistral-7B-v0.1	T4 GPU	0.004048

To integrate performance outcomes into a deployment-oriented comparison, a cost-per-correct-answer metric was calculated, aligning with recent production-readiness analyses [50]. As shown in Table 4.4, GPT-3.5-Turbo delivers a legally accurate output at approximately USD 0.000527, compared with USD 0.000900 for Mistral-7B-v0.1 on L4, a 71% overhead.

Table 4 Cost Per Correct Answer

Model	Cost per Correct Answer (USD)
GPT-3.5-Turbo	**0.000527**
Mistral-7B-v0.1 (L4 GPU)	0.000900

The overall trade-off relationship may be interpreted as follows: GPT-3.5-Turbo dominates on speed, cost, and accuracy, placing it in the optimal region for public deployment; Mistral-7B-v0.1 remains competitive only where data-governance requirements such as data residency under Sri Lanka's Personal Data Protection Act of 2022 outweigh performance differences. Importantly, the open-source model's viability depends heavily on appropriate GPU selection. Lower-end hardware (e.g., T4 GPU) resulted in prohibitively high latency and cost, rendering such configurations unsuitable for operational use.

Statistical tests support these interpretations. A z-test on accuracy yielded p = 0.176 indicating that accuracy differences, while meaningful, are not statistically significant at the 95% confidence level. Conversely, independent t-tests on latency ($p = 1.91\times10^{-31}$) and cost ($p = 1.53\times10^{-20}$) confirm that GPT-3.5-Turbo is significantly faster and more economical. The results therefore validate that latency and cost advantages are systemic rather than incidental.

Collectively, these findings highlight clear deployment implications. GPT-3.5-Turbo provides the strongest balance of responsiveness, legal correctness, and operating cost, making it well-suited for interactive, public-facing legal information services where user experience and throughput are critical. This aligns with guidance that real-world LLM suitability depends on both output quality and performance efficiency [47]. In contrast, Mistral-7B-v0.1, although slower and slightly less accurate, remains an attractive option in institutional contexts that prioritize data residency, privacy, and local infrastructure control, where these trade-offs may be preferable to external API dependency [51]. These results further confirm that in applied legal-AI environments, infrastructure configuration is as consequential as model selection, since hardware characteristics exert a nonlinear influence on latency and operational cost.

V. DISCUSSION

➢ *Implications for Legal Practice and Access to Justice*

The findings of this study demonstrate that retrieval-augmented LLMs can meaningfully support access to legal information within Sri Lanka's civil-law context. Both system configurations using GPT-3.5-Turbo and Mistral-7B-v0.1 successfully interpreted user queries and generated coherent, procedurally relevant responses grounded in Sri Lankan civil law. This capability directly addresses the persistent gap in public legal literacy, especially among citizens with limited access to legal professionals or financial means for early consultation.

In Sri Lanka, structural barriers continue to restrict legal accessibility, particularly in rural and low-income areas. Automating the delivery of procedural guidance for common matters such as divorce, maintenance, property disputes, and fundamental-rights claims could help citizens better understand available legal remedies before approaching traditional institutions such as the Legal Aid Commission. Such technology can also reduce strain on overburdened legal-aid infrastructure by handling initial informational queries at scale.

The comparative system evaluation highlights clear deployment trade-offs. GPT-3.5-Turbo's superior accuracy and faster latency suggest strong suitability for public-facing, interactive tools requiring reliability and trust. However, reliance on foreign cloud services and subscription-based pricing limits feasibility for large-scale, long-term government adoption.

By contrast, Mistral-7B-v0.1, while slower and slightly less accurate, offers substantial benefits in data sovereignty and cost-controlled, on-premise deployment. These strengths align with environments handling sensitive details such as government agencies, legal-aid centres, or NGOs where privacy and infrastructure control are critical. Its performance remains adequate for asynchronous information delivery or internal legal knowledge workflows.

Beyond public-facing legal guidance, such systems can also contribute to enhanced legal research efficiency. They can assist paralegals, junior practitioners, and law students by offering rapid access to procedural explanations, thereby complementing rather than replacing professional reasoning. This aligns with global legal-tech trends positioning AI as a supportive reference companion, not a decision-making authority.

Successful deployment will require continued expert supervision and distinctions between information and formal legal advice. While grounded responses reduce hallucination risks, occasional inaccuracies remain possible. Therefore, clear output disclaimers, regular accuracy audits, and continuous dataset updates are essential to preserve user protection and institutional trust.

Overall, this research demonstrates that RAG-based LLMs can advance Sri Lanka's access-to-justice goals by democratizing legal knowledge. Proprietary and open-source configurations each present viable, but contextually distinct paths toward scalable legal-information services, provided they are deployed with appropriate governance, oversight, and ethical safeguards.

➢ *Limitations*

This study evaluated a functional prototype rather than a production-grade legal AI system, and several limitations define its current scope.

First, the domain coverage was intentionally constrained to common civil-law scenarios (e.g., family law and fundamental rights). As a result, the system does not currently address important areas such as criminal law, commercial disputes, property-title complexities, labour rights, or detailed statutory interpretation. Future expansion should broaden domain diversity to reflect real-world caseload distribution.

Second, legal content reflects the state of the law at the point of dataset creation and lacks automated update mechanisms. Without periodic refresh routines or dynamic retrieval from authoritative sources, accuracy may degrade over time as statutes or procedures evolve.

Third, the system presently supports English only. Given that Sinhala and Tamil are the dominant languages of legal interaction in Sri Lanka, the prototype's accessibility is limited to a subset of the population. Implementing multilingual retrieval and generation capabilities would be critical for equitable public deployment.

Fourth, the performance of the open-source model is hardware-dependent. Testing across GPU configurations (T4, L4, A100) revealed variations in latency and operational cost, indicating that real-world adoption requires careful infrastructure optimization. Additionally, the cost analysis excluded passive standby consumption, an important factor in always-available services.

The dataset size and evaluation scope, while adequate for controlled comparison, limit statistical generalizability. A larger benchmark covering more varied phrasing and nuanced legal situations would strengthen future analytical confidence.

Finally, the current provenance mechanisms link responses at the scenario level rather than providing statute- or section-specific citations. Enhanced metadata granularity would improve transparency, auditability, and trustworthiness for professional legal environments.

These constraints reflect the practical boundaries of a research prototype rather than shortcomings of the underlying approach. Each limitation presents a clear direction for future refinement toward a more comprehensive, multilingual, and operationally robust legal-AI solution for Sri Lanka.

➢ *Ethical Considerations*

Operating within a high-stakes sector, the system's design prioritized user protection, transparency, and responsible data governance. The assistant is explicitly positioned as a legal information tool rather than a provider of legal advice, and response templates include clear disclaimers and guidance to consult qualified legal professionals when necessary. This aligns with internationally recognized principles for Trustworthy AI, which emphasize prevention of harm, human oversight, and accountability throughout the lifecycle of AI deployments. The European Commission's guidelines further stress the importance of transparency, respect for human autonomy, and robustness features embedded here through provenance-linked retrieval, clear communication of limitations, and traceable system outputs [52].

Data-ethics considerations were central to corpus development. All scenario records are hypothetical and anonymized, preventing the exposure of personal or confidential information. No user data were retained during evaluation, supporting compliance with the Personal Data Protection Act No. 9 of 2022 and general privacy-by-design principles. The availability of a fully local deployment pathway via the open-source model further reduces risks associated with cross-border data processing.

Ethical deployment must also account for digital inclusion. The present English-only implementation may inadvertently reinforce linguistic inequities in legal access. Future work must incorporate Sinhala and Tamil support, usability testing, and culturally appropriate interface design to ensure equitable impact.

Sustained oversight is required to manage risk. Human-in-the-loop review, regular expert audits, and monitoring for bias or performance drift are essential governance practices to maintain trust and uphold public-interest objectives.

Collectively, the ethical framework of this project rests on three pillars, 1) Role limitation and disclaimers to avoid unauthorized legal advice, 2) Privacy-protective data governance aligned with Sri Lankan and international standards, 3) Commitment to inclusivity and transparency in public-facing deployment.

By embedding these principles in both design and future implementation planning, this research demonstrates a pathway for responsible legal-AI adoption that protects users while advancing the national goal of equitable access to justice.

VI. CONCLUSION

A. Summary of Findings

This research demonstrates that Retrieval-Augmented Generation (RAG) architectures can provide a viable technical foundation for AI-driven legal information systems within Sri Lanka’s civil-law context. The study performed a systematic comparative evaluation of two contrasting LLM backends GPT-3.5-Turbo (proprietary) and Mistral-7B-v0.1 (open-source) across three deployment-critical metrics: accuracy, latency, and cost. These dimensions align with guidance from prior work emphasizing practical feasibility and responsible AI adoption in specialized legal domains [7][32].

Quantitatively, GPT-3.5-Turbo exhibited superior performance across all primary metrics. It achieved 92.5% accuracy compared to Mistral-7B-v0.1’s 82.5%, delivered a substantially faster average response time (4.17 seconds vs. 15.64 seconds on an L4 GPU), and incurred the lowest mean per-query expense (USD 0.000487 vs. USD 0.000742). When normalized by accuracy, GPT-3.5-Turbo further demonstrated a 71% lower cost per correct response, confirming superior cost-efficiency for reliable outputs. Statistical analysis indicated that while the observed accuracy gap, approximately 10 percentage points, did not reach significance due to the modest sample size, both latency and cost advantages were highly significant ($p < 0.001$), reinforcing their practical importance.

Despite its lower performance in high-throughput settings, Mistral-7B-v0.1 demonstrated credible and legally coherent output, maintaining full data residency and operational autonomy when deployed locally. This deployment model directly aligns with privacy-sensitive legal environments, supporting institutional needs for compliance with national data-governance regulations and the capacity for offline or controlled-access operation. These characteristics make Mistral-7B-v0.1 especially fit for internal knowledge retrieval services such as those in legal-aid offices, universities, and government agencies where instantaneous interaction is not a primary requirement.

The findings also underscore that infrastructure plays a decisive role in open-source LLM performance. Latency varied considerably across GPU configurations (T4, L4, A100), indicating that practical viability depends not solely on model capability but on hardware optimisation and deployment resources. In contrast, GPT-3.5-Turbo benefits from industrial-scale optimization inherent in API-based delivery, ensuring consistent performance even under restricted local compute environments.

From a methodological standpoint, this study demonstrated the feasibility of developing a Sri Lanka–specific RAG pipeline using a curated expert-validated corpus. The system successfully generated contextually grounded procedural guidance and reduced hallucination risk through enforced provenance anchoring. Importantly, the evaluation affirms that small, domain-focused datasets can still enable high retrieval precision, which is a promising outcome for jurisdictions with limited machine-readable legal corpora.

Practically, these results provide evidence that responsible deployment of retrieval-augmented LLMs could meaningfully strengthen public access to justice. By helping users understand procedural pathways before engaging state institutions or legal counsel, such systems could alleviate informational burdens in the legal ecosystem. The differentiated performance profiles further suggest that a hybrid adoption strategy leveraging proprietary models for public-facing, real-time support and open-source systems for secure institutional use may offer the most sustainable path for legal-AI adoption in developing contexts.

➢ *In Conclusion, the Research Establishes that,*

- GPT-3.5-Turbo is best positioned for interactive, high-accuracy, real-time legal-information services.
- Mistral-7B-v0.1 is better aligned to deployments requiring privacy, localization, and infrastructure control.

Both models prove capable of producing legally relevant and procedurally accurate responses when supported by a rigorously constructed RAG pipeline. These outcomes validate the core research objectives and build a strong foundation for advancing toward multilingual, jurisdictionadaptable, and continuously updated AI legal-assistance systems tailored to the Sri Lankan context.

B. *Contributions*

This research offers several interconnected contributions to the field of legal informatics and the application of large language models (LLMs) in access-to-justice initiatives, particularly within low-resource and under-digitised jurisdictions. Collectively, these contributions establish a foundational framework for the development of accurate, transparent, and context-appropriate AI legal-information systems in Sri Lanka and similar environments.

First, the study delivers one of the earliest documented implementations of a retrieval-augmented generation (RAG) system tailored to the Sri Lankan civil-law domain. By designing a compact yet representative corpus of expert-validated legal scenarios and embedding it within a vector-based retrieval pipeline, the research demonstrates the feasibility of generating grounded procedural guidance using modest computational resources. This working prototype serves as a reference architecture for future extensions and real-world deployments in the justice sector.

Second, the study introduces a reproducible comparative evaluation framework for legal-domain LLMs. Through standardized benchmarking of accuracy, latency, and cost, including statistical significance, testing the framework enables fair and evidence-based comparison between proprietary and open-source model configurations. This fills a methodological gap in current scholarship, where evaluations often lack controlled experimental design and operational performance metrics relevant to public-facing systems.

Third, the project contributes the creation of a localized civil-law corpus designed specifically for retrieval-augmented systems. The dataset reflects Sri Lankan procedural realities in areas such as family law and fundamental-rights litigation, and is annotated to support provenance tracking and error analysis. While compact, it is structured for scalability into multilingual, statute-linked, or case-law-integrated datasets, addressing the scarcity of machine-readable legal content in Sri Lanka.

Fourth, the research operationalizes responsible-AI principles directly within system design. Safeguards including strict separation from formal legal advice, provenance-aware prompting, and context-constrained generation align with ethical best practice and national data-governance principles, including the Personal Data Protection Act No. 9 of 2022. This provides a replicable model for ethical AI deployment in high-stakes legal contexts where public trust is critical.

Finally, the study provides policy-relevant evidence to inform technology adoption strategies in the justice sector. The findings demonstrate that GPT-3.5-Turbo offers superior performance for citizen-facing applications requiring rapid and accurate guidance, while Mistral-7B-v0.1 supports localized and privacy-preserving deployments. These insights help institutional decision-makers evaluate trade-offs between performance, cost, data sovereignty, and long-term operational control, core considerations for sustainable AI governance in emerging legal-tech ecosystems.

Together, these contributions demonstrate that responsible, localized legal-AI innovation is technically feasible and socio-legally aligned, even in environments with limited digital infrastructure. The research therefore advances both the academic understanding and the practical readiness of retrieval-augmented systems as tools for strengthening access to justice in Sri Lanka.

C. *Future Research Directions*

While this study demonstrates the feasibility and value of retrieval-augmented LLMs for delivering civil-law information in Sri Lanka, several promising avenues remain for advancing both capability and practical deployment. These opportunities span technical development, linguistic inclusivity, evaluation rigor, and institutional integration, each addressing limitations identified earlier.

A primary direction for future enhancement is broader corpus expansion. The current dataset captures common citizen-facing civil-law matters, but does not yet include areas such as criminal law, commercial transactions, labour law, or administrative appeals. Expanding coverage to these domains would significantly improve the utility of the system for a wider variety of legal needs. Furthermore, integrating primary legal sources including statutory provisions, case law passages, and formal procedural rules would enable citation-level provenance and adherence to professional legal research standards. Semi-automated ingestion from authoritative repositories could also support continuous legal updates, reducing the risk of outdated guidance as laws evolve.

A second major direction is multilingual and culturally contextual deployment. At present, the system operates exclusively in English. For broad public accessibility and alignment with Sri Lanka’s constitutional language obligations, future versions should incorporate Sinhala and Tamil through multilingual fine-tuning or high-quality neural translation pipelines. This would ensure that legal information delivery does not unintentionally reinforce linguistic inequities in access to justice.

Infrastructure optimization represents another key opportunity. Although this study examined per-query inference costs, it did not incorporate standby compute consumption required for real-time system availability, an important factor for public-facing deployments. Future work should explore dynamic scaling strategies such as auto switching GPU allocation, model quantization, or serverless hosting to reduce idle-time expenditure while preserving responsiveness.

Evaluation methodology can also be strengthened. Manual assessment using a fixed test-set, while reliable for a research benchmark, limits statistical generalization. A future continuous evaluation framework leveraging larger datasets, automated fact-verification, and periodic expert audits would provide more rigorous monitoring of accuracy, procedural faithfulness, and model safety over time.

Finally, practical deployment will require additional interaction design and governance mechanisms. Extending the system into a multi-turn conversational assistant capable of clarifying user context would improve its effectiveness in real-world interactions. Institutional partnerships such as with the Legal Aid Commission, courts, or universities could transform the prototype into an operational tool embedded within existing justice-support ecosystems. These collaborations would also enable oversight procedures that ensure transparency, accountability, and social acceptance.

Collectively, these future research directions present a clear roadmap for evolving the current prototype into a robust, nationally deployable legal-AI platform. By expanding legal-domain coverage, incorporating multilingual access, strengthening evaluation and infrastructure, and aligning deployment with justice-sector institutions, next-generation systems could meaningfully accelerate equal access to legal information across Sri Lanka.

REFERENCES

[1]. "Legal framework in Sri Lanka," *aplawjapan*, Jun. 2024 [Accessed: 20 January 2025]

[2]. "The Constitution of the Democratic Socialist Republic of Sri Lanka,". [Online]. Available: https://parliament.lk/files/pdf/constitution.pdf

[3]. Jayasinghe, I. Dilshani, and I. Weerasekara, "Provide Equal Access to Justice and Enhance Legal Awareness Through the Use of ICT," *ResearchGate*, Nov. 26, 2021. https://www.researchgate.net/publication/369358154_Provide_Equal_Access_to_Justice_and_Enhance_Legal_Awareness_Through_the_Use_of_ICT [Accessed: 08 June 2025]

[4]. W. X. Zhao *et al.*, "A survey of large language models," *arXiv.org*, Mar. 31, 2023. https://arxiv.org/abs/2303.18223 [Accessed: 19 February 2025]

[5]. P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks," *arXiv.org*, May 22, 2020. https://arxiv.org/abs/2005.11401 [Accessed: 08 June 2025]

[6]. Dasanayake, C.G. (2024) 'Evaluating the Use of Artificial Intelligence for an Effective Justice System in Sri Lanka'. Available: https://ir.kdu.ac.lk/bitstream/handle/345/7608/LAWJ%20Vol4%20Iss2_2.pdf [Accessed: 07 June 2025]

[7]. S. Liu, Z. Li, R. Ma, H. Zhao, and M. Du, "ContractEval: Benchmarking LLMs for Clause-Level legal risk identification in commercial contracts," *arXiv.org*, Aug. 05, 2025. https://arxiv.org/abs/2508.03080 [Accessed: 02 July 2025]

[8]. J. Wang *et al.*, "Legal evalutions and challenges of large language models," *arXiv.org*, Nov. 15, 2024. https://arxiv.org/abs/2411.10137 [Accessed: 17 August 2025]

[9]. S. A. Thorat and V. Jadhav, "A review on implementation issues of rule-based chatbot systems," *SSRN Electronic Journal*, Jan. 2020, doi: 10.2139/ssrn.3567047 [Accessed: 08 June 2025]

[10]. S. K. Ojha, A. Kumar, T. Bhole, S. Naaz, and CSE Department Sharda University Greater Noida, "Rule-based A.I. chatbot," *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY*, vol. 11, no. 5, pp. 639–640, 2024, [Online]. Available: https://ijirt.org/publishedpaper/IJIRT168361_PAPER.pdf [Accessed: 13 June 2025]

[11]. J. Williamson, "The rise of AI in legal practice: Opportunities, challenges, & ethical considerations," Mar. 21, 2025. https://ctlj.colorado.edu/?p=1297 [Accessed: 17 June 2025]

[12]. O. Isaac and N. Johnson, "The Use of Chatbots in Providing Free Legal Guidance: Benefits and Limitations," *Researchgate.net*. [Online]. Available: https://www.researchgate.net/publication/388179232_The_Use_of_Chatbots_in_Providing_Free_Legal_Guidance_Benefits_and_Limitations. [Accessed: 22 June 2025].

[13]. J. Cui *et al.*, "ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases," Jun. 2023, doi: 10.48550/arxiv.2306.16092. [Accessed: 13 June 2025]

[14]. J. Cui *et al.*, "Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model," *arXiv.org*, Jun. 28, 2023. https://arxiv.org/abs/2306.16092 [Accessed: 13 June 2025]

[15]. L. Bull, "Legal advice chatbot to go head-to-head with living lawyers," *Irish Legal News*, Aug. 03, 2017. [Online]. Available: https://www.irishlegal.com/articles/legal-advice-chatbot-to-go-head-to-head-with-living-lawyers [Accessed: 07 June 2025]

[16]. V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, "Hallucination-Free? Assessing the reliability of leading AI legal research tools," *arXiv.org*, May 30, 2024. https://arxiv.org/abs/2405.20362 [Accessed: 07 June 2025]

[17]. Surani, D. Ho, "AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries | Stanford HAI." https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries [Accessed: 04 September 2025]

[18]. Vassilopoulos, S. Titah, "Can AI chatbots improve access to legal information? A case study on the inhibitors of AI chatbots' implementation in the context of JuridiQC." [Online]. Available: https://biblos.hec.ca/biblio/memoires/vassilopoulos_athena_m2022.pdf [Accessed: 07 June 2025]

[19]. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *Information*, vol. 13, no. 1, p. 41, Jan. 2022, doi: 10.3390/info13010041. [Accessed: 06 June 2025]

[20]. D. Necz, "Rules over words: Regulation of chatbots in the legal market and ethical considerations," *Hungarian Journal of Legal Studies*, vol. 64, no. 3,

pp. 472–485, Jun. 2024, doi: 10.1556/2052.2023.00472. [Accessed: 26 June 2025]

[21]. J. Xue, Y.-C. Wang, C. Wei, X. Liu, J. Woo, and C. -c. J. Kuo, “Bias and Fairness in Chatbots: An Overview,” *arXiv.org*, Sep. 16, 2023. https://arxiv.org/abs/2309.08836 [Accessed: 19 April 2025]

[22]. Palfreyman, “Minimizing legal risks of AI-Powered chatbots,” *Harris Beach Murtha*, Mar. 11, 2025. https://www.harrisbeachmurtha.com/insights/minimizing-legal-risks-of-ai-powered-chatbots/ [Accessed: 19 April 2025]

[23]. V. Karpukhin *et al.*, “Dense passage retrieval for Open-Domain question answering,” *arXiv.org*, Apr. 10, 2020. https://arxiv.org/abs/2004.04906 [Accessed: 08 June 2025]

[24]. Izacard and É. Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering,” *ACL Anthology*, Apr. 2021, doi: 10.18653/v1/2021.eacl-main.74. [Accessed: 22 June 2025]

[25]. N. Pipitone and G. H. Alami, “LegalBench-RAG: a benchmark for Retrieval-Augmented Generation in the legal domain,” *arXiv.org*, Aug. 19, 2024. https://arxiv.org/abs/2408.10343 [Accessed: 13 August 2025]

[26]. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, “Evaluation of Retrieval-Augmented Generation: A survey,” *arXiv.org*, May 13, 2024. https://arxiv.org/abs/2405.07437v1 [Accessed: 16 June 2025]

[27]. F. Petroni*et al.*, “KILT: a Benchmark for Knowledge Intensive Language Tasks,” *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, [Online]. Available: https://aclanthology.org/2021.naacl-main.200.pdf [Accessed: 08 June 2025]

[28]. P. Ahadian and Q. Guan, “A survey on hallucination in large language and foundation models,” *Preprints.org*, May 2025, doi: 10.20944/preprints202504.1236.v2. [Accessed: 17 April 2025]

[29]. Rau 1 *et al.*, “BERGEN: A benchmarking library for Retrieval-Augmented Generation,” *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7640–7663, Nov. 2024, [Online]. Available: https://aclanthology.org/2024.findings-emnlp.449.pdf [Accessed: 03 July 2025]

[30]. Z. Fei *et al.*, “LawBench: Benchmarking legal knowledge of large language models,” *arXiv.org*, Sep. 28, 2023. https://arxiv.org/abs/2309.16289 [Accessed: 03 July 2025]

[31]. Laurent, “LLM Benchmarks in Life Sciences: Comprehensive Overview,” *IntuitionLabs*, Nov. 07, 2025. https://intuitionlabs.ai/articles/large-language-model-benchmarks-life-sciences-overview

[32]. M. S. Y. Alassan, J. L. Espejel, M. Bouhandi, W. Dahhane, and E. H. Ettifouri, “Comparison of Open-Source and Proprietary LLMs for Machine Reading Comprehension: A Practical Analysis for Industrial applications,” *arXiv.org*, Jun. 19, 2024. https://arxiv.org/abs/2406.13713v2 [Accessed: 13 August 2025]

[33]. “LawNet - American Institute for Sri Lankan Studies,” *American Institute for Sri Lankan Studies*. https://www.aisls.org/lawnet/“Search judgments and orders – court of appeal.” https://courtofappeal.lk/?page_id=1191 [Accessed: 03 July 2025]

[34]. “Search judgments and orders – court of appeal.” https://courtofappeal.lk/?page_id=1191 [Accessed: 03 July 2025]

[35]. “Department of Government Printing.” https://documents.gov.lk/

[36]. “Home,” *Srilanka Law*. https://www.srilankalaw.lk/

[37]. “Sri Lankan Legal Materials,” *Commonlii*. https://www.commonlii.org/lk/

[38]. “VLex | Sri Lanka.” https://vlex.com/coverage/sri-lanka

[39]. WorldJusticeProject, “Paths followed by people in Sri Lanka to deal with their everyday justice problems, summarizing the incidence of legal problems, respondents’ legal capability, access to sources of help, problem status, assessment of the resolution process, and problem impact.,” survey, 2017. [Online]. Available: https://worldjusticeproject.org/sites/default/files/documents/Access-to-Justice-2019-SriLanka.pdf [Accessed: 03 July 2025]

[40]. W. N. S. Perera, A. M. Perera, S. Hulathduwa, and P. Paranitharan, “Artificial intelligence-driven digitization of legal system in Sri Lanka - A challenging approach,” *Sri Lanka Journal of Forensic Medicine Science & Law*, vol. 16, no. 1, pp. 51–57, Jun. 2025, doi: 10.4038/sljfmsl.v16i1.8040. [Accessed: 22 June 2025]

[41]. Östling *et al.*, “The Cambridge Law Corpus: a dataset for Legal AI research,” *The Cambridge Law Corpus*, 2023, doi: 10.17863/CAM.100221. [Accessed: 17 June 2025]

[42]. N. Guha *et al.*, “LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models,” *arXiv.org*, Aug. 20, 2023. https://arxiv.org/abs/2308.11462 [Accessed: 07 June 2025]

[43]. S. H. Wang *et al.*, “ACORD: an Expert-Annotated Retrieval Dataset for Legal Contract Drafting,” *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, vol. 1, Jan. 2025, doi: 10.18653/v1/2025.acl-long.1206. [Accessed: 07 June 2025]

[44]. “Chroma is the open-source search and retrieval database for AI applications,” *Chroma*. https://www.trychroma.com/

[45]. Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models: A survey,” *arXiv.org*, Dec. 18, 2023. https://arxiv.org/abs/2312.10997 [Accessed: 08 July 2025]

[46]. Q. Jiang *et al.*, “Mistral 7B,” *arXiv.org*, Oct. 10, 2023. https://arxiv.org/abs/2310.06825 [Accessed: 10 August 2025]

[47]. Bronsdon, "RAG vs Traditional LLMs: Key Differences," *Galileo AI*, Nov. 19, 2024. https://galileo.ai/blog/comparing-rag-and-traditional-llms-which-suits-your-project [Accessed: 09 April 2025]

[48]. "Pricing," *OpenAI*. https://platform.openai.com/docs/pricing

[49]. "Colab paid services pricing." https://colab.research.google.com/signup

[50]. M. H. Erol, B. El, M. Suzgun, M. Yuksekgonul, and J. Zou, "Cost-of-Pass: an economic framework for evaluating language models," *arXiv.org*, Apr. 17, 2025. https://arxiv.org/abs/2504.13359 [Accessed: 20 July 2025]

[51]. "LLM Comparison: key concepts & best practices | NExLa," *Nexla*, Sep. 15, 2025. https://nexla.com/ai-readiness/llm-comparison [Accessed: 23 Sep 2025]

[52]. Publications Office of the European Union, "Ethics guidelines for trustworthy AI.," *Publications Office of the EU*, 2019. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 [Accessed: 06 April 2025]