ERU Engineering Research Unit University of Moratuwa Sri Lanka

Hybrid Approach for Information Retrieval in Sri Lankan Legal Domain

Niduni Kasige
dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
niduni.22@cse.mrt.ac.lk

Nishan Kavinda
dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
nishan.22@cse.mrt.ac.lk

Kisara Kodithuwakku
dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
kisara.22@cse.mrt.ac.lk

Nisansa de Silva
dept. of Computer Science & Engineering
University of Moratuwa
Sri Lanka
NisansaDdS@cse.mrt.ac.lk

Keywords— legal information, hybrid retrieval, semantic search, Sinhala information retrieval

I. Introduction

Information retrieval in the legal domain has become a significant research area due to the specific nature of the language in the legal domain [1]. This complexity is further exacerbated in contexts such as Sri Lanka, where English literacy rates among individuals aged 15 and above are approximately 22%, posing a significant barrier to comprehending intricate legal documentation [2]. Current solutions present a trade-off: traditional keyword searches require technical vocabulary that lay users lack, whereas generic LLMs offer conversational ease but frequently 'hallucinate' or misrepresent local statutes [3]. To address this, we propose a hybrid system combining precise document retrieval with a domain-adapted LLM [4]. Our approach fine-tunes Gemma-3-4B-IT [5] on Sri Lankan legal data [6], combining exact statutory retrieval with simplified summaries for lay audiences.

II. Literature Review

The complexity of legal language creates challenges for effective information retrieval due to their specialized terminology and structured nature [1]. Conventional keyword-based methods such as Boolean search and TF-IDF overlook the contextual subtleties of legal discourse, underscoring the need for semantically aware retrieval systems. Retrieval-Augmented Generation integrates traditional retrieval with generative language models, grounding outputs in authoritative sources to enhance reliability in tasks such as statute interpretation and legal summarization [4]. Hybrid retrieval systems combining keyword and semantic search, further improve contextual accuracy and scalability by bridging symbolic and neural representations. Extractive summarization complements retrieval by distilling key insights from lengthy legal documents, expediting research and comprehension [1][7]. A key ongoing challenge lies in extending these capabilities to multilingual and low-resource settings such as Sinhala, where jurisdiction-specific adaptation and localized retrieval remain crucial for improving legal accessibility and inclusiveness [8][9].

III. Materials and Methods

The development strategy of the proposed system consists of key steps from data acquisition to Sinhala query handling.

A. Data Acquisition

The initial data was sourced from the Sri Lanka legal Document Datasets [8], a JSON collection of Sri Lankan Acts organized by metadata and sections, enabling efficient document chunking and embedding generation.

B. Document Processing and Retrieval

Document cleaning removes extra spaces, special characters, and common patterns. Texts are chunked based on document structure (e.g., constitution chunked by chapters, then paragraph-based segments). Vector embeddings are generated using Legal-BERT [10], and BM25 scores [11] are computed for each document. The retrieval system operates using BM25 keyword search and FAISS indices for bills, acts, gazettes, and the constitution.

C. Extractive Summarization

The system generates key highlights from retrieved documents using extractive summarization. Cosine similarity between sentence pairs is calculated, and sentences with higher similarity to multiple sentences are selected as key highlights.

D. Fine-tuned LLM

To meet domain requirements under tight compute limits, we fine-tuned the Gemma-3-4B-IT model [5] using the `Unsloth` framework. The training data was sourced from a legal-conversations dataset [6], formatted as multi-turn dialogues. To optimize efficiency, we employed Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) [12] (`r=8, α=8`), specifically targeting attention, language, and MLP modules. Training utilized the 8-bit `AdamW` optimizer (lr=2e-4) with gradient accumulation to ensure stable domain adaptation. The model was deployed through Hugging Face endpoints.

Evaluation used 30 legal queries across domains with the following metrics: ROUGE scores, semantic similarity

(Legal-BERT), BERTScore F1, factual consistency, citation accuracy, context relevance, response length, and time. Both systems utilize 245,864 chunks of legal documents across four categories.

E. Sinhala Query Handling

The final stage handles Sinhala queries by fine-tuning an existing embedding model to suit legal texts. Although Sin-BERT, RoBERTa, and LaBSE support Sinhala, they lack domain-specific understanding, while Legal-BERT [10] does not cover Sinhala legal texts. Transfer learning (contrastive learning) with a LoRA adapter was applied to Sin-BERT-large (XLM-RoBERTa architecture). The Sri Lankan constitution was used to build the dataset, with Google Document AI for accurate text extraction [9]. Sentence chunks were prepared, and the model was evaluated using the Sentence-Transformers Embedding Similarity Evaluator.

IV. Results and Discussion

A. Qualitative Retrieval Example

{ **"query"**: "What is the punishment for using counterfeit current coin as genuine?",
"retrieved_documents": [
- "*Penal Code of Sri Lanka - Section 236*: Delivery of coin, possessed with knowledge that it is counterfeit...",
- "*Penal Code of Sri Lanka - Section 478B*: Using as genuine, forged or counterfeit currency notes or bank ...",
- "*Central Bank of Sri Lanka Act, No. 16 of 2023 - Section 37*: Sole authority to issue currency notes and coins...",
- "*Prevention of Money Laundering Act, No. 5 of 2006 - Section 3*: Offence of money laundering. Any person who.."],
"generated_answer": "According to the Central Bank of Sri Lanka Act, No. 16 of 2023, the Central Bank is the sole authority for issuing currency, and using forged currency is a serious offense. Furthermore, under the Prevention of Money Laundering Act, No. 5 of 2006, assets derived from counterfeiting can be forfeited. Regarding the specific punishment, the Penal Code of Sri Lanka, Section 478B states that using as genuine any forged or counterfeit currency note is punishable with imprisonment of up to twenty years and a fine."}

Table 1 presents performance metrics on English test data, demonstrating clear advantages of domain-specific finetuning across all evaluation dimensions.

The fine-tuned model shows strong overall performance, with a 63.0% relative improvement in ROUGE-1, 5.4% in semantic similarity, and major improvements in factual consistency (209.6%). This reflects better use of statutory language and reduced hallucinations.

While Gemini (gemini-2.5-flash) achieves higher citation accuracy (0.5833 vs 0.4500) and faster responses (3.0s vs 11.85s), the fine-tuned model provides more comprehensive, factually grounded answers essential for accuracy-critical legal applications.

While fine-tuned Gemma3 4B focuses on generating legal responses, the fine-tuned Sin-BERT model is an embedding model for measuring semantic similarity between legal texts. It achieved an accuracy of 0.57 in identifying contextual similarity, while suggesting further improvements on dissimilar text identification.

TABLE 1. COMPREHENSIVE METRICS COMPARISON ACROSS ALL DIMENSIONS

METRIC	FINE-TUNED	GEMINI
ROUGE-1	**0.3373**	0.2069
ROUGE-2	**0.1973**	0.0608
ROUGE-L	**0.2975**	0.1426
Semantic Similarity	**0.5875**	0.5574
BERTScore F1	**0.8669**	0.8341
Factual Consistency	**0.4466**	0.1442
Citation Accuracy	0.4500	**0.5833**
Context Relevance	**0.3195**	0.2958
Response Time (s)	11.85	**3.00**

V. Conclusion and Future Work

This study presents a system for accessing legal information using a hybrid keyword and semantic search approach, enhanced with extractive summaries for easier comprehension. It also extends support for Sinhala, enabling semantic understanding of Sinhala legal texts.

Future work will focus on improving the fine-tuned Sinhala embedding model and developing a pipeline for Sinhala legal text tokenization and domain-specific stop word identification.

References

[1] Ajay Mukund, S., & Easwarakumar, K. S. (2025). Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation. Symmetry, 17(5), 633.

[2] Abayasekara, A. (2018, April 23). Building a more English-literate Sri Lanka: The need to combat inequities. Talking Economics (IPS Blog). https://www.ips.lk/talkingeconomics/2018/04/23/building-a-more-english-literate-sri-lanka-the-need-to-combat-inequities/

[3] Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. T. (2023). Do language models know when they're hallucinating references? arXiv:2305.18248.

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv:2005.11401.

[5] Gemma Team. (2024). Gemma: Open models based on Gemini research and technology. arXiv:2403.08295.

[6] Nishan726. (2025). Sri Lankan Legal Conversations. Hugging Face Datasets. https://huggingface.co/datasets/Nishan726/sri-lankan-legal-conversations

[7] Panchal, D., Gole, A., Narute, V., & Joshi, R. (2025). LawPal: A retrieval-augmented generation-based system for enhanced legal accessibility in India. arXiv:2502.16573.

[8] Senaratna, N. I. (2025). Sri Lanka Document Datasets: A large-scale, multilingual resource for law, news, and policy. arXiv:2510.04124.

[9] Jayatilleke, N., & de Silva, N. (2025). SiDiaC: Sinhala Diachronic Corpus. arXiv:2509.17912.

[10] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT. arXiv:2010.02559.

[11] Robertson, S. E., & Zaragoza, H. (2009). BM25 and beyond. Foundations and Trends in Information Retrieval.

[12] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA. arXiv:2106.09685.