SMALL LANGUAGE MODELS FOR SRI LANKAN LEGAL APPLICATIONS

Project ID - 25-26J-240

Project Proposal Report

Abiramy.T - IT22049322

Thuvaraga.A - IT22030412

Niruthika.E – IT22322326

Mathusigan.S - IT22117032

Bachelor of Science (Hons) Degree in Information Technology
Specializing in Information Technology

Department of Information Technology

Sri Lanka institute of information technology Sri Lanka (SLIIT)

July 2025

PREDICTS LIKELY LEGAL JUDGMENTS BY ANALYZING PAST CASES IN CRIMINAL LAW

Project ID - 25-26J-240

Project Proposal Report

Abiramy.T - IT22049322

Bachelor of Science (Hons) Degree in Information Technology
Specializing in Information Technology

Department of Information Technology

Sri Lanka institute of information technology Sri Lanka (SLIIT)

July 2025

DECLARATION

I declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously publish or written by another person expect where the acknowledgement is made in the text.

Name	Student ID	Signature
T. Abiramy	IT22049322	T. Abiramy

The supervisor/s should certify the proposal report with the following declaration.

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

..... 2025/9/18.

Signature of the supervisor: Date:

Dr. Prasanna Sumathipala

....

Signature of the Co-Supervisor Date:

Ms.Karthiga Rajendran

ABSTRACT

This research study proposes the use of an intelligent web application to predict Sri Lankan criminal case verdicts in addressing the inefficiencies and unfairness of the nation's hybrid legal system, where Roman Dutch, English common, and local laws coexist. Sri Lanka's criminal justice system is unable to provide timely and accessible justice due to human and other resource limitations, lack of digitization, and procedural delay, especially in Penal Code offenses like theft, assault, and murder. To provide accurate, useful predictions, the system employs a Legal-BERT model that has been trained on condensed datasets of Sri Lankan criminal cases. With a highly scalable and user-friendly system architecture, the legal system is made for lawyers and law students who work within the legal profession. The frontend application is developed with React.js and Next.js, and the backend is developed with FastAPI. The test will be focused on accuracy, response time, and impartiality with the aim of improving legal decision-making, relieving congestion in backlogs, and democratizing legal professionals' access to justice.

Keywords

Criminal Case Prediction, Sri Lanka Legal System, Legal-BERT, BERT, Transformer model, Legal Judgement Prediction, Deep Learning

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVITIONS

GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
LJP	Legal Judgment Prediction
NLP	Natural Language Processing
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
DPO	Direct Preference Optimization
NMSP	Neural Machine Translation with Shared Parameters
SLM	Small Language Model
TLM	Transformer-based Language Model

1. INTRODUCTION

The legal system in Sri Lanka is governed by a wide range of laws across multiple domains, such as property, family, labour, and criminal law.Though these laws offer advanced legal advice, their complexity, disjointed organisation, and technical language makes them challenging for non-specialists to understand and use, and even a few lawyers.

Small Language Models (SLMs) tailored to local legal contexts can now be created thanks to developments in Artificial Intelligence (AI), especially in Natural Language Processing (NLP). Four elements are the main components of this study on small language models for legal applications in Sri Lanka:

1. Fine-tuning an SLMs to provide step-by-step end-to-end guidance for solving user problems in Sri Lankan Property Law and Family Law.

2. Fine-tuning an SLMs to review and analyze risks in Sri Lankan deed documents.

3. Developing a law recommendation system for lawyers, with a special focus on Labor and Employment Law.

4. Finetuning an SLMs to predict likely legal judgments by analyzing past case data, enabling users to understand relevant precedents and potential outcomes.

Using old case data examined by a fine-tuned Legal-BERT model with the help of OpenAI's summarisation feature, my research component focuses on developing a user-friendly web application that can assist in predicting the judgement of criminal cases in Sri Lanka. It provides straightforward, intelligible predictions that are easy for anyone to comprehend. This system is intended to facilitate comprehension of prior court laws, particularly about common offences like homicide, assault, and theft that are covered by Sri Lanka's distinctive blend of local, English, and Roman Dutch laws. This system, designed for lawyers and law students, intends to reduce delays, improve the legal system, and increase access to justice.

1.1 Background & Literature survey

Sri Lanka's legal web application presents a complicated combination of Roman Dutch law, English common law, and local traditions, making significant challenges in the criminal justice sector governed by agencies like the Penal Code of Sri Lanka. Inefficiencies occur from procedural delays, limited digitization of case data in references such as the Sri Lanka Law Report documents, and manual precedent retrieval, resulting in case backlogs, high costs, and unequal access to justice. These problems are specifically critical in low resource environments, where large language models (LLMs) like GPT-3 impracticable due to computational needs, energy

consumption, and privacy risks [10]. Small language models (SLMs) propose a hopeful option for edge AI deployments, allowing on device intelligence with efficient architectures and quantization methods [10]. This research study component creates a web application for predicting criminal case judgments using a fine-tuned Legal-BERT model trained on past case data, focusing on Penal Code offences to improve efficiency and democratize justice in Sri Lanka.

1.1.1 Similar Works

- Aralimatti et al. [10] present the Shakti SLM series (100M, 250M, 500M parameters) for domain-specific advantage AI, including legal applications, making use of architectures like GPT-3 and LLaMA while including Rotary Positional Embeddings (RoPE) and Grouped Query Attention (GQA) for efficiency, with benchmarks on jobs like MMLU and Hellaswag, and domain-specific evaluations in healthcare, finance, and legal.

- Jayasinghe et al. [13] present domain-specific additional models for legal case succeeding party prediction using U.S. Supreme Court data, using RoBERTa sentence embeddings and transformer encoders to gain 75.75% accuracy with critical sentence annotation, categorizing methods into political or social science based, linguistics-based, or legal domain-based attributes.

- Ariai et al. [15] survey NLP studies in law, emphasizing transformer architectures for judgment prediction (LJP), such as MPBFN for multi-subtask dependences, and models like Legal-BERT fine-tuned on legal corpora, with discussions on jobs including Document Summarization, Named Entity Recognition, Question Answering, Argument Mining, Text Classification, and Judgement Prediction.

- Nguyen et al. [21] detail Transformer-based methods in COLIEE 2021, utilizing BERT variants for legal entailment and question responding, with changes like NFSP and NMSP leveraging multilanguage similar translations, gaining competitive performance in tasks applying case law retrieval, entailment, and statute law question responding.

- Greco and Tagarelli [3] summary Transformer-based language models (TLMs) in legal AI, emphasizing BERTology's role in case retrieval, entailment, and prediction, with variants like Legal-BERT and extensions for multilanguage contexts, discussing applications in agreement review, legal analysis, and result prediction, while highlighting fast progress since the first Transformer model.

1.1.2 Technologies

- Transformer architectures with self-attention instruments, as in BERT's bidirectional fine-tuning for legal text learning, including methods like ALBERT, ELECTRA, and BART [21] [3]

- Shakti SLMs with optimized architectures utilizing RoPE, GQA, and quantization (int4/int8/int5) for edge devices, allowing efficient inference on constrained hardware like Raspberry Pi [10].

- RNN-based (GRU/LSTM) and transformer encoder architectures, combining BiLSTM-CRF for ordering tasks like named entity recognition and critical sentence identification [13].

- NN architectures like Siamese BERT for similarity, DistilBERT with extended attention for extended legal texts, and models such as T5, XLNet, and GPT-NeoX containing segment level replication and prefix-based objectives [3] [15].

- Multilanguage models like NMSP/NFSP leveraging similar translations, with methods such as Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Reinforcement Learning from Human Feedback (RLHF) for alignment and bias relief [21][10].

1.1.3 Limitations in Existing Work

- Resource Constraints: LLMs like GPT-3 need necessary GPUs, limiting edge deployment in low resource environments; SLMs mitigate this but may compromise accurateness [10] [3].

- Data Absence and Bias: Lawful datasets are usually unstructured or biased, risking unjust results; multilanguage gaps remain for languages like Sinhala/Tamil [21] [15].

- Localization Issues: Models trained on Western corpora (e.g., U.S. cases) overlook mixed systems like Sri Lanka's, with limited transformation for fine performances [13] [3].

- Interpretability and Ethics: Unclear TLM logic and biases in benchmarks like BBQ/ToxiGen delay trust; moral considerations include privacy and toxicity [10] [15].

- Predictive Accuracy: Support on structured components falls in complicated legal texts; lengthy documents and cross-references pose challenges [13] [21].

1.2 Research Gap

While important advances have been made in Transformer-based models for legal NLP jobs, several critical gaps continue, specifically in resource-constrained, multilanguage, and mixed legal backgrounds like Sri Lanka's criminal judge system. Existing works, such as those on SLMs for edge AI [10] and domain-specific models for case prediction [13], present efficiency in Western-centric contexts but fall short in handling localized challenges. Surveys underline the demand for more interpretable, bias-mitigated models tailored to various legal fields [15][3]. Yet, applications in producing countries remain restricted due to data absence, increased computational needs of LLMs, and poor multilanguage support [21]. This study is important to bridge these gaps by implementing a lightweight, AI web application for criminal judgment prediction, leveraging Legal-BERT and OpenAI summarization on Sri Lankan precedents.

The following significant gaps highlight the need for this research:

- Insufficient Localization to Legal Systems: Models like Legal-BERT and Shakti SLMs are mainly trained on common law corpora (e.g., U.S. Supreme Court data), managing Sri Lanka's unique mix of Roman Dutch, English common, and local laws, leading to inaccuracies in slight variations of the Penal Code [13] [15] [3].

- Resource and Deployment Restrictions in Low Resource Environments: LLMs such as GPT-3 needs comprehensive GPUs, casuing them unfit for edge devices in Sri Lanka; SLMs provide possible but lack of exact transformations for legal edge AI, including quantization and device privacy [10][21].

- Data, Bias, and Multilangage Shortage: Legal datasets are usually unstructured or limited (e.g., via benchmarks like BBQ/ToxiGen), with limited support for Sinhala/Tamil alongside English, leading to unfair predictions in various contexts [15] [3] [10].

- Gaps in Predictive Accurateness and Interpretability for Judgment Tasks: Current methods rely on structured components or Western precedents, failing to manage long, complicated Sri Lankan case documents; multilanguage models like NFSP/NMSP provides a guarantee results but its require integration with summarization for actionable results [13] [21].

- Moral and Realistic Challenges in Real-World Applications: The Unclear reason, toxicity risks, and moral issues in bias relief (e.g., via RLHF/DPO) slow adoption. there is a demand for a system that continusly improves access to justice while providing justice [10] [15].

This research component project addresses these gaps by organizing localized criminal case report data, fine-tuning SLMs for edge deployment, and integrating bias mitigation methods, eventually providing an affordable tool to reduce procedural waits and democratize lawful predictions in Sri Lanka.

1.3 Research Problem

The critical issue of inefficient and unavailable criminal case judgment prediction within Sri Lanka's difficult mixed system, which is defined by a hybrid framework of English common law, Roman Dutch law, and indigenous law, is addressed by this research component. The existing criminal justice sector, managed by the Penal Code of Sri Lanka, suffers from procedural holds, limited digitization of case data from sources like the Sri Lanka Law Reports, and relies on manual precedent retrieval, leading to large case backlogs, high litigation expenses, and not equal access to justice. Existing AI-based system solutions, such as large language models (LLMs) like GPT-3 and transformer-based models like Legal-BERT, are ill-suited for this context due to their increased algorithmic needs and lack of localization to Sri Lanka's unique legal and multilanguage (Sinhala/Tamil/English) environment [10][15][3]. Likewise, the absence of structured, digitized legal case data and the existence of preferences in available corpora interfere with accurate predictions [21] [13].

The exact problem this research addresses is the absence of a lightweight, localized web application that can indicate criminal case judgment based on past case reports, customized to Sri Lanka's low-resource environment.

Main issues include:

- **Inaccessibility of Case data**: Legal practitioners and the public people struggle to get and analyze appropriate criminal case report data due to its unstructured qualities and manually processing requirements [15].
- **Resource Limitations**: The computational needs of LLMs and the absence of optimized small language models (SLMs) for edge deployment limitation scalability in Sri Lanka's infrastructure [10].
- **Multilanguage and Cultural Misalignment**: Existing models, trained on Western corpora, fail to account for Sri Lanka's multilanguage legal texts and mixed legal nuances, causing incorrect or limited outputs [3] [13].
- **Lack of Predictive Implements**: There is no available web solution that utilizes a fine-tuned Legal-BERT model to provide actionable judgment predictions for Penal Code offenses (e.g., theft, assault, homicide) [21].

This research component solves these problems by implementing an web application that utilizing Legal-BERT trained on summarized Sri Lankan criminal case report data for predictions, and ensures compatibility and bias mitigation, thereby improving efficiency and equal access to justice.

2. OBJECTIVES

2.1 Main Objectives

The main goal of this research component is to implement an extremely user-friendly and accessible web application that accurately predicts criminal case judgments in Sri Lanka's typical and complicated judicial system. By addressing the long running issues of inefficiency, inaccessibility, and unfairness that affect the criminal justice system, which is controlled by a complicated combination of Roman Dutch law, English common law, and local legal practices, this initiative aims to change the way justice is delivered. By harnessing the strength of advanced artificial intelligence tailored to local needs, the project seeks to empower lawyers and law students with a solution that delivers accurate, timely, and reliable predictions of case results, specifically for overall crimes under the Penal Code of Sri Lanka, such as theft, assault, and homicide. The goal of this initiative is to close the gap between advanced technology and practical implementation in an environment with limited resources, thereby paving the way for a more open, effective, and equal legal system. By decreasing procedural holds, lowering the financial and time constraints of litigation, and democratizing access to legal knowledge, the application will be a transformative tool that helps create a more equitable society where individuals from all backgrounds can more easily navigate and comprehend the criminal justice system.

2.2 Specific Objectives

1. To ensure that lawyers and law students may easily use a web interface that allows them to enter case facts and determine expected judgments using a refined Legal-BERT model.

2. To improve Legal-BERT on trained datasets of criminal cases in Sri Lanka so that case predictions that are appropriate for the hybrid legal system may be accurately predicted.

3. To use OpenAI summary abilities to produce detailed, straightforward case outcome predictions, making the tool useful and understandable for users.

4. To make sure the application can function very well in Sri Lanka's resource-constrained environment while protecting user privacy and cutting down on legal process delays, it should be optimized for low-resource edge devices.

3. METHODOLOGY

This research assumes a multi-stage, holistic process to design, develop and test a web application for criminal case judgment prediction in Sri Lanka's criminal justice system. The approach draws on principles of natural Language processing (NLP) and artificial intelligence (AI) with special modifications to address the unique challenges of a hybrid legal system that combines Roman Dutch law, English common law, and indigenous traditions. Taking a leaf from domain-specific fine-tuning techniques for small language models (SLMs) introduced in Aralimatti et al [10], and transformer-based techniques for legal text processing in Nguyen et al. [21] note that the method prioritizes efficiency, localization, and ethical factors. It integrates data curation, model adaptation, and system architecture design, and rigorous validation to ensure the expedience of the application. The methodology is iterative with the possibility for enhancement at every intermediate evaluation and prioritizes deployment to minimize computational needs to a bare minimum while maximizing access to lawyers and law students.

3.1 Data Collection and Preparation

Data collection is the foundation of any AI-driven legal application, and for Sri Lanka's criminal justice system, it is particularly challenging to find suitable, high-quality information due to a lack of digitization and multilingual competency. This phase begins with the systematic collection of historical criminal case information on previous violations of the Penal Code of Sri Lanka (Ordinance No. 2 of 1883), such as theft, assault, murder, and related offenses. Sources are mainly publicly available repositories like the Sri Lanka Law Reports, Supreme Court rulings published in government gazettes, and Ministry of Justice digitized collections. To supplement them, anonymized legal database data will be added, with ensuring observation of data under data privacy legislation, like Sri Lanka's Personal Data Protection Act.

The collection process will be guided by relevance, recency, and diversity criteria to provide a broad range of case types across different courts (e.g., Magistrate's Courts, High Courts, and Court of Appeal) and language differences (Sinhala, Tamil, English). Up to 500-1000 cases will be addressed first, with the balance in representation to prevent biases, as underscored in bias reduction considerations in Aralimatti et al [10] and Ariai et al. [15]. Ethical principles are foremost: all the data will be anonymized to remove identifiers, and approvals will be secured where necessary to accommodate responsible AI standards.

After collection, preparation of data has some subprocesses that transform raw, unstructured legal texts into a machine-readable format for model training. Preprocessing will include text cleaning to remove artifacts like formatting bugs, footnotes, and repeated headers, using regular expressions and NLTK libraries structuring of data is necessary; every case would be dissected into pieces: Case facts, quoted legal provisions (e.g., specific sections of the Penal Code), arguments between

the petitioner and the defendant, and final judgment. This is similar to the domain-specific additional models of Jayasinghe et al. [13], in which sentence annotation for critical sentences enhanced prediction performance.

Multilingual management is an innovation first here, bridging the gap identified in Greco and Tagarelli [3] for non-English law texts. Sinhala and Tamil texts will be translated to English with parallel principles or multilingual models like NMSP/NFSP from Nguyen et al. [21], with semantic fidelity. Data augmentation techniques will be used to augment the dataset, including synonym substitution, back-translation, and synthetic case generation using rule-based problems, to address data scarcity, a common flaw in legal NLP as identified by Ariai et al. [15]. The dataset will subsequently be split into training (70%), validation (15%), and testing (15%) sets, with stratification to maintain balance across case outcomes (e.g., conviction and acquittal).

3.2 Model Development

The spirit of the application is constructing models that make successful judgment prediction from legal texts. Following the inspiration from the Shakti SLM series in Aralimatti et al. [10], that focuses on optimal architecture for domain-specific tasks, this phase has its focus on fine-tuning Legal-BERT for prediction. Legal-BERT is selected because of its bidirectional context understanding, crucial for achieving the minimal meanings in Sri Lankan legal precedents, elaborated in Greco and Tagarelli [3].

Fine-tuning begins from a pre-trained Legal-BERT model to work with legal inputs, fine-tuned on the ready, abridged legal principle using supervised techniques like those in Nguyen et al. [21]. The objective is training the model for case outcome direct prediction (i.e., classification as conviction, acquittal, or specific sentencing).Learning rate of hyperparameters (2e-5), batch size (16), and epochs (3-5) will be adjusted using grid search with early stopping to prevent overfitting. To meet constraints, quantization-aware training will be used, reducing model size up to 4x without degrading accuracy, as in Aralimatti et al. [10].

Bias mitigation is integrated here, taking advantage of responsible AI mechanisms in Aralimatti et al [10], such as diverse sampling of data and debiasing prompts to avoid cultural or gender biases common in legal data (e.g., via benchmarks like BBQ and ToxiGen).

3.3 System Design and Integration

System design focuses on the integration of the created model into a consistent web application that includes the created model into a friendly platform for lawyers and law students. The interface, built on React.js and Next.js, will be an easy-to-use form

where one can input detailed crime information—like various kinds of crimes (for instance, homicide and robbery), number of witnesses, crime scene, and brief description of what happened, with dynamic, server-side rendering experience with Tailwind CSS styling so it is accessible and usable, in line with the goals of enhanced legal decision-making in resource-constrained environments.

Integration contains producing the workflow: user provided form input is passed through the fine-tuned Legal-BERT model in order to generate predictions, which include retrieving references to similar past cases from the training dataset based on semantic similarity. The output will display a shortened summary of the probable judgment (e.g., "Likely conviction under Penal Code Section 296 with a sentence of 5-7 years imprisonment, considering mitigating factors like evidence of witnesses"), along with key sources from previous case overviews (e.g., highest 3-5 relevant precedents with short quotes and references). For localizability, the system will incorporate Sri Lanka-specific legal ontologies like Penal Code section mappings and fusion legal concepts (Roman Dutch, English common, and indigenous customs), borrowing ideas from domain-specific embeddings in supporting research. Privacy is safeguarded by on-device processing, where possible by avoiding cloud dependency, and by rectifying data authority problems in developing markets.

The backend, caused by Fast API in Python, will process API requests, run Legal-BERT predictions, and perform reference retrieval with high-performance asynchronous processing, ensuring smooth communication with the React.js and Next.js frontend. This technology stack is supportive of real-time data processing and virtual security functionalities, improving the overall efficiency and reliability of the application.

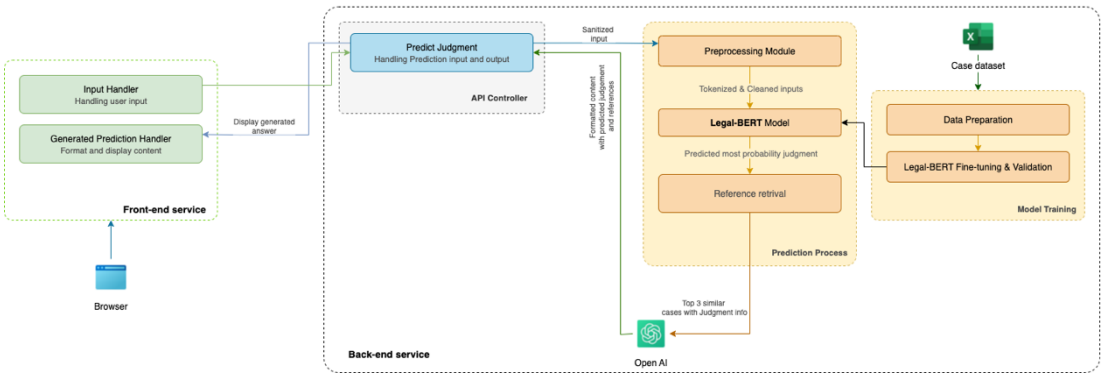

Figure 1: System architecture diagram

3.4 Evaluation and Refinement

Evaluation is necessary to confirm the effectiveness, precision, and equity of the application. A multi-faceted framework shall be used with focus on human evaluation and score-based metrics. Quantitative measurement shall include retrieval performance by precision@K and recall@K (targeting >75%), summarization quality by ROUGE and BLEU scores (>0.5), and general prediction accuracy by F1-score

over a held-out test set (>70%). Bias audits shall use metrics like demographic parity and equalized odds to detect and mitigate issues in multilingual predictions.

Qualitative testing will include human tests with 20-30 stakeholders (e.g., lawyers, judges, and citizens) via surveys and interviews to test usability, interpretability, and perceived trustworthiness. Participants will provide scores (e.g., 1-5 scale) on clarity, relevance, and confidence in predictions with a minimum average target score of 4.0. Realistic simulations will test the application on unseen cases, measuring latency (less than 5 seconds on edge devices) and noisy input resistance, with user feedback employed in further refinement.

Improvement will be incremental: based on findings from evaluation, models will be retrained with enhanced data, prompts optimized for better summarization, and biases mitigated by approaches like Direct Preference Optimization (DPO). Discussion of ethical review will ensure adherence to Sri Lankan legal requirements, with the outcome a deployable prototype that advances AI for justice in low resource settings.

4. PERSONAL AND FACILITIES

As the proposed web application for criminal case judgment prediction in Sri Lanka is naturally connected to the expert knowledge domain of criminal judgment, predictions made by the system must be valid, appropriate, and moral. With legal understanding of criminal cases requiring expertise that exceeds the technical competency of the project team, legal expert input has been integrated into the project.

4.1 Facilities and Resources

The success of this project depends upon the availability of technical and legal resources. Main facilities are:

- Legal Data Access: Official government legal websites (e.g., Ministry of Justice website, LawNet), online legal databases, and published legal reports (e.g., SLR and NR).

- Computing Infrastructure: Training systems for models with GPUs, cloud infrastructure for storing and accessing data at scale, and secure servers to host the deployed system.

- Software Tools: Python, FastAPI React.js for web development, Hugging Face Transformers for NLP BERT model fine-tuning.

Together, these expert inputs and resources will make sure that the proposed system is technologically sound and legally sound, meeting both the twin necessities of encouraging AI research as well as addressing outstanding gaps in Sri Lankan Labor and Employment Law accessibility.

4.2 Technical Guidance

To complement the legal expertise, technical supervision will be provided by academic mentors specializing in Artificial Intelligence (AI) and Knowledge Systems:

- Dr. Prasanna Sumathipala (Supervisor) - Expert in Artificial Intelligence and Natural Language Processing, responsible for guiding the fine-tuning of the TS model and the integration of NLP techniques.

- Ms.Karthiga Rajendran (Co-supervisor) - Specialist in Information Retrieval and Knowledge-Based Systems, providing support in implementing Retrieval-Augmented Generation (RAG) and ensuring efficient data handling

Their combined expertise will ensure the seamless integration of advanced Al methodologies into the legal decision support framework.

5. GANTT CHART

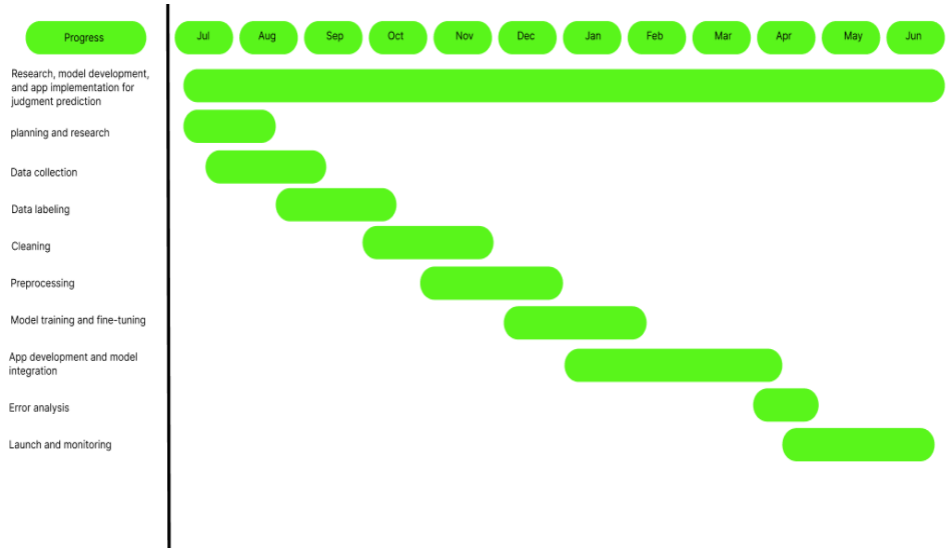

Figure 2: Gantt Chart

6. PROJECT REQUIREMENTS

6.1 Functional requirements

1) Case Input Interface: The web application must have the ability to input case data (e.g., crime type, circumstances, law provisions) to make a judgment prediction request.

2) Judgment Prediction: The system must generate a straightforward prediction of the likely result of the case (e.g., conviction, acquittal, sentencing details) based on the Legal-BERT fine-tuned model, with probabilistic confidence ratings and justification.

3) Multilanguage Support: The system must accept input and give output in Sinhala, Tamil, or English, with the availability of real-time translation to enable access regardless of language preference.

4) User Feedback Mechanism: There should be a user feedback option in the system for prediction accuracy and usability so that iterative improvements can be made.

6.2 Non-functional Requirements

1) Performance: To allow for real-time use, the system must be able to answer questions and make predictions on edge devices (such as smartphones and Raspberry Pi) in less than five seconds.

2) Scalability: In the first deployment, the application must be able to accommodate up to 100 users at once. It should also be scalable as its use grows by leveraging cloud and edge hybrid architecture.

3) Reliability: Using a held-out dataset of criminal cases from Sri Lanka, the system must show a minimum prediction and retrieval accuracy of 70% while maintaining 99% uptime on edge devices.

4) Security: To preserve user privacy and adhere to Sri Lankan data protection laws, data must be encrypted (AES-256) both in transit and at rest, as well as on-device.

5) Usability: The interface must have a straightforward design, clear instructions, and an intuitive user experience for non-technical users (such as citizens or lawyers). It must also receive at least 80% of the possible points in user testing.

6.3 Technology Selection

1) Natural Language Processing: Legal-BERT fine-tuned would be utilized for legal text understanding and result prediction, leveraging its bidirectional context feature for exact matching.

2) Summarization: OpenAI API (e.g., GPT-3.5-turbo) would be utilized for advanced summarization to generate human-readable judgment predictions, integrated using secured API calls.

3) Frontend Development: React.js and Next.js will be selected to create a dynamic, server-side rendered frontend with quick loading times, SEO support, and compatibility with both mobile and desktop devices, styled with Tailwind CSS for a responsive, modern design.

4) Backend Development: Fast API will be selected to create a high-performance, asynchronous backend to handle API request.

7. SYSTEM DIAGRAM

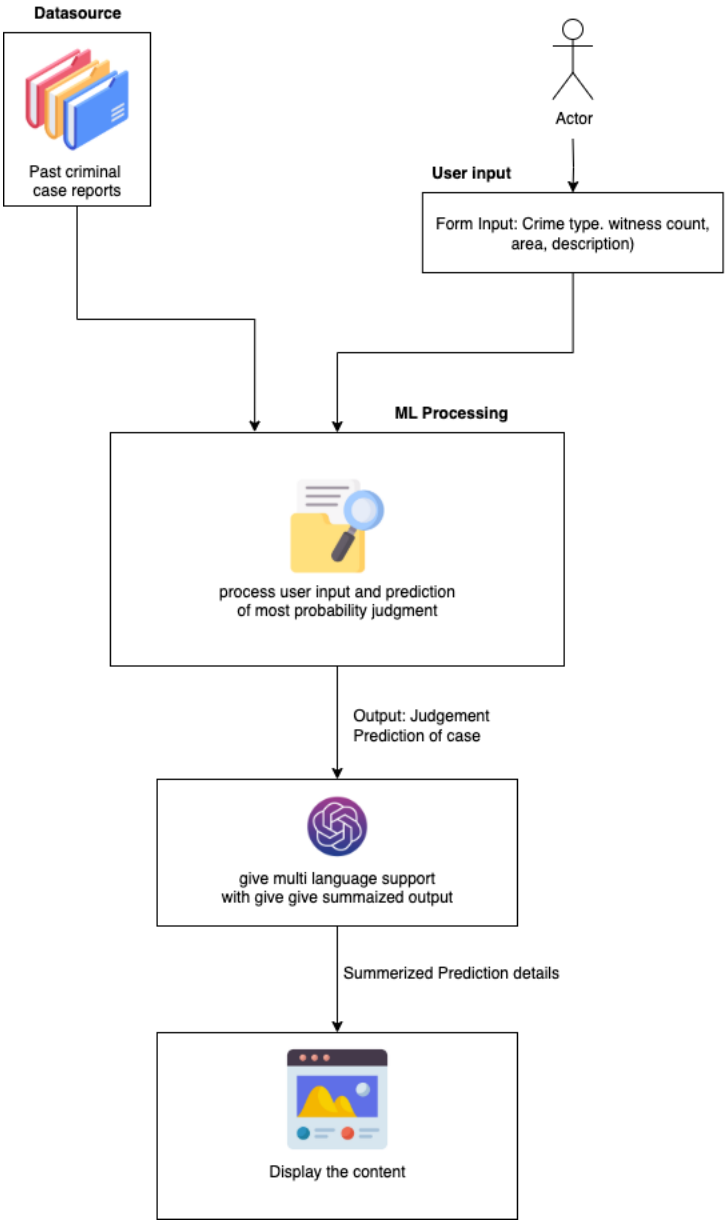

Figure 3 : System Diagram

8. CONCLUSION

This proposal tells of a pioneering effort to overcome Sri Lanka's inefficiencies and injustices of the criminal justice system by creating a new web application for criminal case judgment prediction. The project fills the pressing need for easy access to localized legal instruments in Sri Lanka's Roman Dutch, English common, and local hybrid legal system, under which procedural conditions, limited computerization, and lack of resources have stretched justice provision for centuries. Using a Legal-BERT model optimized on abridged sets of Sri Lankan criminal cases and OpenAI's summarization feature for brief outputs, the proposed application promises to revolutionize legal practitioners' and the public's access to Penal Code-related cases such as theft, assault, and murder.

The solution provides a strong, iterative method of data collection, model construction, and system integration, optimized for edge deployment in Sri Lanka's low-resource environment. The research gaps identified, localization problem, resource limitations, lack of data availability, predictive accuracy, and ethical concerns are dealt with head-on, confirming the system to satisfy the country's multilingual needs (Sinhala, Tamil, English) and ethical requirements. The three-tier modular structure-based system design prioritizes scalability, privacy, and ease of use, making it a proper solution to minimize litigation expenditure, minimize backlogs of cases, and facilitate empowerment of access to justice.

With its successful deployment, this project will not only make legal judgment-making more effective but also empower oppressed communities by giving them an accessible portal to sort through complex legal precedents. The strict evaluation framework, such as precision, recall, ROUGE scores, and bias audits, will guarantee that the application is reliable and fair, paving the way towards further enhancement and wider usage. This research presents a significant step toward integrating AI into Sri Lanka's judicial system and proposes a model that could encourage similar efforts in other developing countries facing such challenges. Moving further into the future, the prospect of continued development of this system using more data and advanced AI methods holds the promise of a more equitable and accessible judicial system.

9. BUDGET AND BUDGET JUSTIFICATION

Table 1: Budget and Budget Justification

Item	**Description**	**Cost (LKR)**
Cloud Computing Resources.	Covers the cost of cloud-based infrastructure (e.g., AWS or GCP) for training the Legal-BERT model and utilizing OpenAI API for summarization, ensuring efficient processing and scalability during development and testing.	10,000.00 LKR
Advertisement	Online and offline promotions to raise awareness of the system.	10,000.00 LKR
Data Digitization	Funds the conversion of past criminal case records from Sri Lanka Law Reports and other sources into digital format to build the dataset.	2000.00 LKR
OpenAI API Subscription	Scanning physical legal documents for dataset creation.	5000.00 LKR
Documentation Printing	Printing research reports, manuals, and related documents.	2000.00 LKR
Internet Cost	Internet access for research, cloud usage, and collaboration	5000.00 LKR
Total		**34,000.00 LKR**

The estimated cost for the web application of Criminal Case Judgment Prediction in Sri Lanka is LKR 34,000.00. This budget covers needed expenses on cloud computing resources to facilitate model training and utilization of the OpenAI API, marketing activities to create awareness among the general public and legal experts, digitization of documented criminal case files, printing materials on user guides and assessment reports, and internet usage for data synchronization and testing. These expenses ensure hassle free development, testing, and deployment of the application with ease, and its implementation and outreach within Sri Lanka's criminal judge system.

10. COMMERCIALIZATION

10.1 Market Opportunity

Sri Lanka, like most developing nations, struggles with adapting to advanced and ever-evolving legal systems, especially that of criminal justice. The proposed web application addresses this shortfall through the application of Natural Language Processing (tuned Legal-BERT) to provide closely related criminal case judgment predictions, complete with references to Penal Code titles, sections, enactment years, and accompanying case scenarios.

Likely Users are:

- Law Firms & Attorneys: Reduce research time and get correct legal precedents.
- Business & HR Departments: Get compliance-related facts at once for employee problems.
- Government Agencies & Legal Aid Offices: Provide accurate, timely responses to the public.
- Courts & Labour Tribunals: Rapid checks of laws and related case situations.
- Colleges & Law Students: Assist learning better with organized examples and direct quotes.
- General Public: Receive plain, accurate legal information without professional hurdles.

Unique Features:

- Provides accurate law citations (section, year, scenario) instead of vague keyword matches.
- Particularly created for Sri Lankan law with the facility to incorporate other jurisdictions.
- Forces integration with third party apps such as Law case management software, 3rd party apps.

10.2 Revenue Model

1) Business Subscriptions (B2B):
    a. Small Plan: Solo lawyers and small firms (limited queries/month).
    b. Professional Plan: Small-sized businesses with additional amenities and higher limits

c. Enterprise Plan: Large firms, government, and universities with unlimited questions, bespoke dashboards, and enhanced.

2) Government & Institutions:

    a. Annual licenses for ministries, labor departments and universities.

3) Freemium (Public Access):

    a. Free Tier: Limited queries with simplified answers.

    b. Paid Upgrade: Complete access, complete citations, printable reports.

4) Add-Ones:

    a. Pre-formatted legal templates, automated compliance checklist of lists, and premium scenario packs.

5) White-Label Offering:

    a. Reselling and implementing technology to organizations on their own systems on a licensed basis.

REFERENCES

[1] R. S. D. Oliveira and E. G. Sperandio Nascimento, "Analysing Similarities between Legal Court Documents Using Natural Language Processing Approaches Based on Transformers," 2024, *SSRN*. doi: 10.2139/ssrn.4944655.

[2] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Exploratory Research in Clinical and Social Pharmacy*, vol. 13, p. 100419, Mar. 2024, doi: 10.1016/j.rcsop.2024.100419.

[3] C. M. Greco and A. Tagarelli, "Bringing order into the realm of Transformer-based language models for artificial intelligence and law," *Artif Intell Law*, vol. 32, no. 4, pp. 863–1010, Dec. 2024, doi: 10.1007/s10506-023-09374-7.

[4] "Bringing order into the realm of Transformer-based language models for artificial intelligence and law | Artificial Intelligence and Law." Accessed: Aug. 16, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s10506-023-09374-7

[5] H. Alberts, A. Ipek, R. Lucas, and P. Wozny, "COLIEE 2020: Legal Information Retrieval and Entailment with Legal Embeddings and Boosting," in *New Frontiers in Artificial Intelligence*, vol. 12758, N. Okazaki, K. Yada, K. Satoh, and K. Mineshima, Eds., in Lecture Notes in Computer Science, vol. 12758. , Cham: Springer International Publishing, 2021, pp. 211–225. doi: 10.1007/978-3-030-79942-7_14.

[6] J. Rabelo, M.-Y. Kim, and R. Goebel, "Combining Similarity and Transformer Methods for Case Law Entailment," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, Montreal QC Canada: ACM, June 2019, pp. 290–296. doi: 10.1145/3322640.3326741.

[7] M. J. Davenport, "Enhancing Legal Document Analysis with Large Language Models: A Structured Approach to Accuracy, Context Preservation, and Risk Mitigation," *OJML*, vol. 15, no. 02, pp. 232–280, 2025, doi: 10.4236/ojml.2025.152016.

[8] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches," *IEEE Access*, vol. 13, pp. 18253–18276, 2025, doi: 10.1109/ACCESS.2025.3533217.

[9] A. Abdallah, B. Piryani, and A. Jatowt, "Exploring the state of the art in legal QA systems," *J Big Data*, vol. 10, no. 1, p. 127, Aug. 2023, doi: 10.1186/s40537-023-00802-8.

[10] R. Aralimatti, S. A. G. Shakhadri, K. KR, and K. B. Angadi, "Fine-Tuning Small Language Models for Domain-Specific AI: An Edge AI Perspective," Mar. 03, 2025, *arXiv*: arXiv:2503.01933. doi: 10.48550/arXiv.2503.01933.

[11] D. Licari and G. Comandè, "ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain," *Computer Law & Security Review*, vol. 52, p. 105908, Apr. 2024, doi: 10.1016/j.clsr.2023.105908.

[12] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, 2021, doi: 10.1016/j.aiopen.2021.06.003.

[13] S. Jayasinghe, L. Rambukkanage, A. Silva, N. de Silva, and A. S. Perera, "Legal Case Winning Party Prediction With Domain Specific Auxiliary Models," 2022.

[14] M.-Y. Kim, J. Rabelo, K. Okeke, and R. Goebel, "Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods," *Rev Socionetwork Strat*, vol. 16, no. 1, pp. 157–174, Apr. 2022, doi: 10.1007/s12626-022-00103-1.

[15] F. Ariai, J. Mackenzie, and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," July 30, 2025, *arXiv*: arXiv:2410.21306. doi: 10.48550/arXiv.2410.21306.

[16] N. Okazaki, K. Yada, K. Satoh, and K. Mineshima, Eds., *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers*, vol. 12758. in Lecture Notes in Computer Science, vol. 12758. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-79942-7.

[17] A. A. Dzaky *et al.*, "Optimization Chatbot Services Based on DNN-Bert for Mental Health of University Students," *JAIC*, vol. 8, no. 1, pp. 13–21, July 2024, doi: 10.30871/jaic.v8i1.7403.

[18] Y. Yin and I. Habernal, "Privacy-Preserving Models for Legal Natural Language Processing," in *Proceedings of the Natural Legal Language Processing Workshop 2022*, 2022, pp. 172–183. doi: 10.18653/v1/2022.nllp-1.14.

[19] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis, "Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer," Nov. 10, 2022, *arXiv*: arXiv:2211.00974. doi: 10.48550/arXiv.2211.00974.

[20] A. Jeet Rawat, S. Ghildiyal, and A. K. Dixit, "Topic modelling of legal documents using NLP and bidirectional encoder representations from transformers," *IJEECS*, vol. 28, no. 3, p. 1749, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1749-1755.

[21] H.-T. Nguyen *et al.*, "Transformer-Based Approaches for Legal Text Processing: JNLP Team - COLIEE 2021," *Rev Socionetwork Strat*, vol. 16, no. 1, pp. 135–155, Apr. 2022, doi: 10.1007/s12626-022-00102-2.