SMALL LANGUAGE MODELS FOR SRI LANKAN LEGAL APPLICATIONS

25-26J-240

Project Proposal Report

Thuvaraga Anantharajah
Mathusigan Senthan
Abiramy Thiyakesan
Niruththika Erambanathan

B.Sc. (Hons) Degree in Information Technology
Specializing in Information Technology

Department of Information Technology
Sri Lanka Institute of Information Technology Sri Lanka

August 2025

DEVELOPING AN ALGORITHM FOR TEMPLATE MATCHING IN DEED DOCUMENTS

25-26J-240

Project Proposal Report

Thuvaraga Anantharajah

B.Sc. (Hons) Degree in Information Technology
Specializing in Information Technology

Department of Information Technology
Sri Lanka Institute of Information Technology
Sri Lanka

August 2025

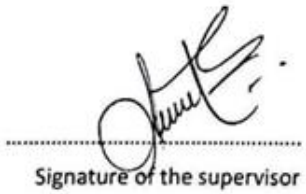
DECLARATION

We declare that this is our 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 published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature
A. Thuvaraga	IT22030412	A. Thuvaraga

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

Name of supervisor: DR.Prasanna Sumathipala



Signature of the supervisor

2025/7/18

Date

ABSTRACT

Deeds are the pillars of transfer of property and inheritance and should be proper and consistent to prevent litigation. Human verification is time-consuming and expensive and liable to miss details in bulky papers or poor scans. Extremely sophisticated AI systems can reduce it to some extent but are too expensive and complicated with restricted resources. We investigate a possible alternative in this research work: small language models (SLMs) with explicit rule-based template matching to identify anomalies in deeds. We process with five common document classes Power of Attorney, Deed of Transfer/Conveyance (Sale Deed), Deed of Gift, Last Will & Testamentary Deed, and Deed of Mortgage and produce two outputs automatically: (1) a concise document summary and (2) a conformity report that highlights each alert with the exact lines where issues exist. We have a four-step pipeline. Step 1: Ingest and validation. We ingest the file, verify format and size and if scanning, conduct OCR and record a quality score. Step 2: Information extraction. Light-weight models and pattern rules are trained on legal texts and pull out major fields (parties, dates, property description, amounts, witnesses, notary details, encumbrances).Step 3: Deed classification and agent routing. A deed-type router classifies and sends the document to the appropriate deed agent. Within each agent, we run a template-matching algorithm that integrates (a) firm rules for mandatory clauses and cross-checks (e.g., attestation, chain of title, acceptance of gift, executor appointment, loan and security terms) with (b) narrow SLM prompts for intent and language consistency (e.g., excessively broad powers, latent conditions, unclear boundaries to property). The algorithm produces issue type, severity, and evidence ranges (page/line). Step 4: Scoring and explanation. Checks are given a score of Low/Medium/High and tabulated as a total conformity score. A confidence measure is calculated from OCR quality and extraction coverage. We tested the method on a mix of clean PDFs and noisy scans. Routing to the right agent plus the agent-internal template checks helped small models work well with narrow prompts and deed-specific rules. The system consistently flagged important problems such as missing attestation details, name mismatches, unclear boundaries, absent acceptance language in gift deeds, overwide powers in powers of attorney, and incomplete release terms in mortgages. Reviewers reported that evidence-linked findings made verification faster and more reliable. The system ran quickly and showed a clear "low confidence" warning when text quality was poor.

Keywords: small language models (SLM), natural language processing (NLP), template matching of deeds, legal-tech, artificial intelligence (AI).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Key Words	Meaning
SLM	Small Language model
API	Application Programming Interface
NLP	Natural Language Processing
AI	Artificial Intelligence

Table 1:List of Abbreviation

1.INTRODUCTION

It produces real-world applied legal-tech products viable for Sri Lankan practice. It has four pillars:

1. Fine-tuning a small language model (SLM), facilitating end-to-end, step-by-step directions in Sri Lankan real estate law and family law.

2. Template-matching verification of deeds (this component): lightweight SLMs plus explicit rule checks that automatically classify deeds, extract fields, and verify conformity to expected deed templates. For five common documents Power of Attorney, Sale Deed, Gift Deed, Last Will, and Mortgage Deed the system produces (a) a short document summary and (b) a conformity report that highlights deviations from template expectations with exact page/line evidence and a confidence indicator.

3. A law-recommendation system for citizens, with an initial specialization in labor and employment.

4. Predictive models that predict probable legal judgments from the past

The deed verification component uses template matching and small language models (SLMs) to review five common deed types in Sri Lanka. It ingests PDFs or scans, extracts key fields, and routes documents to deed-specific agents that apply mandatory template checks and narrow SLM prompts to spot ambiguous wording or missing clauses. Issues are flagged with severity and page/line evidence, producing a summary and a conformity report with confidence scores. Tested on clean and noisy scans, the system reliably caught common errors while keeping review faster, cheaper, and easier to audit.

1.1 BACKGROUND AND LITERATURE REVIEW

Property succession and transfer in Sri Lanka are very dependent on legal documents such as powers of attorney, sale deeds, gift deeds, testamentary deeds, and mortgage deeds. These documents must be accurate, consistent, and legal. In practice, most deeds are lengthy, stamped and signed, scanned and printed. Manual checking is labor-intensive, costly, and bound to omit crucial errors. Big language AI can offer deed review but is expensive to run and involves privacy concerns when handling sensitive land records. A more practical option is using small language models (SLMs) with transparent template-matching checks, in addition to presenting evidence for every finding. The target is to create a "deed verification agent" which determines the deed type, extracts major fields, enforces rules, uses focused SLM reasoning, and finally produces a summary, conformity score, and page/line pointers.

Current work suggests that machine learning can flag up errant text in legal agreements and assist human reviewers with easy-to-understand and simple signals [1], [24]. Recent works have shifted towards clause-level template checking and creating benchmarks to assess the capacity of models to detect absent or ill-formed clauses, which can be useful for training and evaluation of a deed-verification system [2]. System level, other works recommend combining expert rules with retrieval-augmented generation (RAG) and refinement methods to reduce errors and increase trust [3], [4], [22]. These can be applied directly to deeds since the mechanism must point to certain sentences in the document and, if needed, point to statutes or practice notes

Another field of research is into long and complex legal documents. Experiments show that breaking long documents into shorter sections before analysis, so models can handle them more easily [5]. Experiments also show that neural approaches read legal documents better than conventional methods [6]. Legal NLP reviews indicate progress in classification, entity recognition, summarization, and judgment prediction but also indicate problems such as handling very long documents, multilingually, and generating trustworthy, explainable results [7], [8]. For summarization, current work emphasizes

precise, citation-based summaries over generic abstracts, which is precisely what is needed for a deed-verification system [9]

Although judgment prediction is not interested in deeds, techniques from this domain are relevant. Multi-task techniques that correlate related predictions improve accuracy and reliability [10]. Feedback-based methodologies reduce errors, and semantic extraction finds salient facts [11], [12]. These ideas inspire cross-checking processes for deeds, like verification of names and dates across sections and more robust confidence scores. Closer to contracts and deeds, research has shown that language models with added legal knowledge improve clause checking [13]. Altered models such as LEGAL-BERT establish strong baselines for clause-level template matching in legal documents [14], [17]. ACORD datasets provide support for clause ranking and retrieval, which is valuable for detecting the exact text that passes or fails a template test [15], [16]. Assessment instruments like LegalBench and RAG-benchmarking measure retrieval and reasoning quality and can be applied to deed verification [22]

In the wider document verification space, it is noted that AI must be able to explain itself and report uncertainty with prudence [23]. Work in the previous year also suggests small models, as agents and with retrieval and rule-based included, can be cheap and trustworthy for low-resource, privacy-sensitive settings [20]. Legal models that derive outputs based on actual laws and regulations further boost trust, especially in applying the methods to enact them in Sri Lankan deeds [21], [4].

LLM-Used Papers

Paper	Technologies Used	Strengths	Limitations
[2] Contract Eval	LLM prompts; clause-level risk labels; evaluation suite	Standardized benchmark; focuses on risk at clause level; useful metrics	Contracts domain (not deeds); large-model cost; limited multilingual
[3] Reliable Legal AI (modular)	Expert rules RAG; guardrails/refinement; pipeline orchestration	Reduces hallucinations; auditability; explainable workflow	Complex to build; needs curated KB; not deed-specific
[4]/[22] Bridging Legal Knowledge & AI	Vector store; knowledge graph; hierarchical NMF; retrieval + grounding	Strong grounding; better snippet retrieval; interpretable links	KB creation/maintenance overhead; latency; infra heavy
[7] Legal NLP Survey (2024)	Survey of LLM tasks/datasets/models	Broad landscape view; gaps & challenges summarized	No system or code; high-level only
[8] Legal NLP Survey (2023)	Survey across pre-LLM and LLM era	Historical context; taxonomy of tasks	Not implementation-focused; jurisdiction-agnostic

[9] Legal Summarization Survey (2025)	LLM/extractive–abstractive summarizers; factuality checks	Guidance for evidence-linked summaries; eval pitfalls	Survey only; no ready model
[13] Knowledge-Augmented LLM for Contract Risk	Domain knowledge injection; prompted LLM risk flags	Better risk detection with domain hints; transferable idea	Construction contracts focus; knowledge curation effort; compute cost
[15]/[16] ACORD (Clause Retrieval Dataset)	Bi-encoder retrieval; cross-encoder/ LLM re-ranking	Large expert dataset; trains evidence retrieval	Contract clauses (not deeds); mainly English; setup effort
[21] LawGPT (legal LLM, CN)	Legal corpus pretraining; instruction tuning; knowledge-enhanced	Domain gains for QA/reasoning; template for legal tuning	Non-Sri Lankan law; big model footprint; adaptation required

Table 2 :LLM-Used Papers

Agents-Used Papers

Paper	Technologies Used	Strengths	Limitations
[3] Reliable Legal AI (modular)	Expert rules; RAG (vector store/KG); guardrails/refinement; orchestration	Lowers hallucinations; explainable and auditable; modular, domain-grounded	High integration effort; curated KB needed; latency/ops overhead; not deed-specific
[4]/[22] Bridging Legal Knowledge & AI	Retrieval stack: vector store + KG + hierarchical NMF; RAG pipeline	Precise snippet retrieval; interpretable links; stronger grounding	KB/KG build & maintenance cost; heavier infra; multilingual setup needed
[20] SLMs for Agentic AI	Small LMs; tool-use; planners; multi-agent orchestration; quantization	Low cost; fast; on-prem friendly; composable agents	Narrower knowledge than big LLMs; needs careful task decomposition; more engineering glue
[11] Multi-Perspective Bi-Feedback Network	Predict-then-check loop; attention; multi-signal training	Built-in verification reduces errors; robust pattern for 'verify' stage	Requires labeled data; not a deed pipeline; explanations limited without spans

[10] Topological Multi-Task LJP	DAG/graph of dependent sub-tasks; joint training	Captures task dependencies;improves consistency; informs agent step ordering	Case/judgment focus (not deeds); needs well-defined task labels; transfer requires mapping

Table 3:Agents-Used Papers

Small Language Model–Used Paper

Paper	Technologies Used	Strengths	Limitations
[1] Chakrabarti et al. (2019)	Doc2Vec; traditional classifiers (SVM/LogReg); paragraph-level risk scoring	Compute-light; easy to train; explainable features	Limited context understanding; manual features; not deed-specific
[5] Long-length Legal Doc Classification (2019)	Segmenting long docs; BiLSTM/attention; chunk aggregation	Works with long PDFs; small models; simple deployment	Loses cross-segment context; requires careful chunking
[6] Empirical DL for Legal Review (2018)	CNN/compact neural models vs classical baselines	Better than SVMs with enough data; fast inference	Needs labeled data; weaker on nuanced reasoning
[10] Topological LJP (2018)	DAG of dependent subtasks; multi-task learning (non-LLM)	Captures task dependencies; efficient	Case-law focus; mapping to deeds required

[11] Bi-Feedback LJP (2019)	Predict-then-check loop; attention; compact models	Built-in verification; reduces false positives	Needs curated labels; explanations limited without spans
[12] SLJP (2023)	Transformer encoder; semantic extraction; chunked processing	Strong accuracy with moderate size; good for long text	Not pretrained on Sri Lankan deeds; GPU helpful for training
[14] Leveraging BERT for Legal Classification (2021)	BERT-base; domain tuning; long-doc strategies	Solid baseline; adaptable to deeds; moderate compute	512-token limit; needs segment/merge logic
[17] LEGAL-BERT (2020)	Legal-domain BERT variants; domain pretraining	Better legal vocabulary; drop-in backbone for NER/classification	English-centric; still mid-size; local corpus needed
[18] Real-Estate Risk Assessment (2021)	ANP/MCDA scoring; lightweight analytics	Clear weighting of factors; inspires risk aggregation	Non-NLP; no text understanding; domain shift to deeds
[19] XAI for Credit Risk (2025)	Tree/linear models; SHAP/LIME explanations	Transparent scoring; user-friendly justifications	Tabular focus; needs adaptation for text spans
[24] AI in Legal Domain (2019)	Classic ML/NLP pipeline overview	Historical baselines; low compute	Outdated vs transformers; high manual feature work

Table 4:Small Language Model–Used Papers

1.2 RESEARCH GAP

There is active interest in using AI to read legal documents, but Sri Lankan deed documents are an under-served market. Research and the majority of tools focus on contracts or court cases in other countries, and little work is done on the specific checks that notaries and conveyancers make in local deeds. This research aims to fill the following gaps:

1. No public dataset – There is no publicly available, well-annotated collection of Sri Lankan deeds by general categories (transfer, gift, power of attorney, mortgage, testamentary) so there is no way to train or compare models.

2. Local rules missing – Every type of deed has its set of requirements (e.g., acceptance language for gifts, attestation information, chain of title, priority/release in mortgage), but they are not enacted as machine-checkable template rules.

3. Messy scanned layouts – Real actions hold low-resolution images, faded ink, pages, tables, or annexed survey plans tilted, and the majority of research standards cannot detect these.

4. Lack of evidence-linked results – Specialists need page/line evidence for each alert, but most AI systems simply provide labels or scores without offering the exact text span.

5. Use of large models – Much of the current work is based on costly cloud-based LLMs, and the potential of small language models (SLMs) based on template matching on personal machines remains to be completely unlocked.

6. Scarce legal grounding – Outputs are rarely cited with Sri Lankan statutes, circulars, or practice notes, and no standard exists for retrieval of law in deeds.

7. No common evaluation criterion – There is no Sri Lanka-specific taxonomy of issues, severity levels, or evidence-spared annotated datasets, and thus comparison is unjust.

8. No time consciousness – Acts must be compared to the law prevailing on the date of signing, but temporal rule changeability is addressed by few systems.

9. No human-in-the-loop training – Notary corrections and feedback are not detected often enough in order to improve system performance safely.

10. Incomplete workflow – Analysis will generally cover a single task (OCR, NER, or classification), whereas real life requires an end-to-end process: intake → validation → deed-type routing → field extraction → template checks → SLM reasoning → legal retrieval → conformity scoring → report.

11. Privacy and deployment unmet – Much Sri Lankan office work requires on-premises, offline systems with secure storage and audit logs, yet solutions available do not usually do this.

12. Fairness untested – There are few studies that quantify performance across Sinhala, Tamil, or older deed forms, and precision may differ by location or registry.

Deed template matching Tools Comparison

Product	Primary scope	Key tech / features	Strengths	Limitations vs your goals	Fit for Sri Lankan deed risk
Deed Reader Pro	Metes-and-bounds plotting from deed text	OCR → parse bearings/distances → CAD/IntelliCAD export; Windows app; free trial	Very fast plotting; surveyor-friendly outputs; practical desktop tool	No legal risk analysis; no statute grounding; not Sri Lanka-specific; geometry focus	Low

Deed Plotter AI	AI plotting for deeds/easements/leases	OCR + NLP/LLM → JSON → map/plot; high-volume processing	Automates plotting at scale; integrations; modern web tool	No clause/issue risk checks; not deed-type agents; not localized to SL law	Low
V7 Go – Deed Analysis Agent	Title-exam extraction & summarization	Agentic workflow; reads deeds; extracts ownership data, legal descriptions, encumbrances	Speeds title review; agent platform; enterprise support	Black-box LLM stack; cloud/infra heavy; not tailored to SL deed rules; unclear evidence-span granularity	Medium
AiPazz	Legal research (Sri Lanka)	Search Sri Lankan cases/legislation; AI insights	Local database; helpful for research and citations	Not a deed parser or risk engine; no OCR/plotting	Medium (as a legal source, not analyzer)
Your SLM deed-risk agent	Risk analysis for 5 deed types (PoA, Transfer/Sale, Gift, Testamentary, Mortgage)	On-prem SLM + rules + optional RAG; OCR (Si/Ta/En); deed classifier; issue severity + evidence spans + confidence	Evidence-linked, explainable; Sri Lanka rule checklists; privacy-first; lightweight	Needs local dataset/annotations; OCR robustness; future registry/RAG links	High

1.3 RESEARCH PROBLEM

Sri Lankan property transmissions and inheritances are founded on deed documents like transfer/conveyance deeds, gift deeds, powers of attorney, testamentary deeds, and mortgages. They are long, scanned, and complex. Manual checking is laborious, costly, and prone to errors, which causes disputes, fraud, or delays in registration or property transfer. With continually improving AI methods for legal text, much of the published work targets foreign contracts or case law, and some tools are built on large, expensive models that are difficult to deploy in in-house offices [1], [2], [7], [8], [20]. Evidence supports that machine learning can assist reviewers by identifying clauses not conforming to standard patterns, and metrics at the clause level enable performance to be quantified in emphasizing non-conforming text [1], [2]. Strong legal AI is more likely to combine rules with retrieval-augmented generation (RAG) to ground outputs on authoritative evidence and restrict errors [3], [4], [22]. Studies on lengthy legal text indicate the use of text segmentation and powerful architectures, which is essential for deeds longer than the capacities of typical model sizes support [5], [6]. Surveys identify current challenges such as long context, domain shift, and the need for explainable outputs on the basis of cited evidence [7], [8], [9]. Knowledge-infused models already augment contract verification, suggesting a pathway towards expertise-based template-matching tests in deeds [13]. While research concentrates on the aspect that AI must be transparent, calibrated, and auditable, especially for impactful decisions [23].

All notaries, conveyancers, lenders, buyers, and public offices in Sri Lanka depend on accurate, timely examination of deeds. If errors are ignored—such as missing attestation, unclear property borders, missing language of acceptance in gift deeds, or missing mortgage release terms—parties may end up in court, losing money, or with delays. Cloud-based big-model solutions are expense- and privacy-problems-related for sensitive land data and do not suit low-resource settings. There is a clear need for a tool which runs locally, cleans up noisy scans, and produces evidence-linked outputs that can be verified by a human [7], [8], [20], [23].

No such powerful Sri Lanka-focused system exists today that: (i) stores local deed-type rules in machine-readable templates, (ii) analyzes multilingual, scanned documents with robust field extraction, (iii) draws its conclusions via RAG from Sri Lankan legal sources, and (iv) works based on small language models (SLMs) with clear evidence spans and calibrated confidences. There are no public datasets and Sri Lankan deed evaluation criteria (e.g., labels and page/line evidence) available, hence it is difficult to compare and improve fairly [2], [7], [9], [15], [22].
It bridges the gap by developing an explainable, low-cost, SLM-based template-matching Sri Lankan deed verification agent. The agent will classify deed type, retrieve significant fields, conduct deed-specific template verification, use target SLM prompts in ambiguous cases, and anchor outputs with evidence lines and confidence levels. It is designed for private or on-prem installation and tested against a Sri Lanka-specific annotation scheme (template deviations + evidence spans), closing the local gap between general legal NLP advancements and the local deed review practitioners' real requirements [1], [2], [4], [5], [7], [13], [20], [22], [23].

2. OBJECTIVES

This study is part of a Sri Lankan legal AI program using small language models (SLMs).The focus here is Component 2: template-matching deed analysis with an agent algorithm. The objectives below define what we will build and how we will measure success.

2.1 MAIN OBJECTIVES

The main objectives of this project are to develop an on-premises, privacy-preserving deed–template-matching agent capable of automatically classifying Sri Lankan deeds, including sale, lease, and mortgage documents, across Sinhala, Tamil, and English languages. The system aims to accurately extract critical fields such as parties, dates, property details, and key legal clauses from both scanned and digital documents. It will apply deed-specific templates and standardized structures, enhanced by focused small-language-model reasoning, to identify missing, extra, or altered clauses. In addition, the agent will generate concise, evidence-linked template-matching reports and summaries with clear page and line citations to support findings. By integrating multilingual OCR and ensuring local deployment, the solution will maintain high accuracy while safeguarding sensitive legal information and ensuring compliance with privacy requirements.

2.2 SPECIFIC OBJECTIVES

1. Define the Domain : Build a Sri Lankan deed taxonomy (transfer, gift, power of attorney, testamentary, mortgage) and a template taxonomy (required, optional, prohibited clauses) aligned to local practice.

2. Encode Templates : Write machine-readable deed templates per type (e.g., attestation/witness requirements, acceptance for gifts, chain of title, mortgage priority/release, executor appointment).

3. Create a Dataset : Curate a private, de-identified set of deeds with language mix (Si/Ta/En), scan noise, and gold labels for: deed type, extracted fields, clause mappings, deviations, and evidence spans.

4. Build Intake & OCR : Implement validation and OCR with quality scoring; add post-OCR cleanup for mixed scripts, stamps, seals, and handwritten inserts.

5. Train a Deed Classifier : Develop an SLM-based classifier with a target accuracy ≥90% across deed types and languages; report macro-F1 and fairness by language.

6. Develop the Template-Matching Agent : Inside each deed agent, align extracted clauses with standard templates; detect missing, extra, or altered sections; design deviation scoring (Minor/Moderate/Major).

7. Add Legal Grounding (RAG) : Retrieve and cite relevant Sri Lankan statutes/circulars; support time-aware checks (apply the rule valid on the deed date).

8. Link Evidence : Return page/line text spans for every deviation; evaluate with span-overlap metrics (e.g., token F1 / character IoU).

9. Calibrate Confidence : Produce a confidence indicator using OCR quality, extraction coverage, and model probabilities; evaluate with Brier score / ECE.

10. Evaluate End-to-End : Measure classification accuracy, extraction F1, template-matching deviation detection precision/recall/F1, retrieval hit rate, latency (<2 min/deed on modest hardware), and review-time reduction in a small user study.

11. Run Ablations & Robustness Tests : Compare template-only vs SLM-only vs hybrid; test robustness to scan noise and mixed scripts; analyze common failure modes.

12. Deliver a Prototype : Ship a privacy-preserving on-prem app (CLI/Web) that generates a document summary and template-matching deviation report with audit logs, along with documentation and limits.

3. METHODOLOGY

3.1 OVERVIEW (PIPELINE)

Upload & validate check file type/size; detect if scanned; store a document ID.

OCR & clean : if scanned, run OCR (Sinhala/Tamil/English), deskew, denoise, and page-split. Save an OCR quality score (0–1).

Field extraction : extract parties, dates, property identifiers, consideration/amounts, witnesses, notary, encumbrances, plan numbers.

Deed classification : classify into one of five types: Power of Attorney, Transfer/Conveyance (Sale), Gift, Testamentary Deed (Last Will & Testament), Mortgage.

Agent routing : send the file to the correct deed agent.

Template matching (inside agent) : align extracted clauses/fields to the deed-type template; detect missing, extra, or altered clauses; link every deviation to page/line evidence.

Scoring & confidence : assign per-deviation severity (Minor/Moderate/Major), compute an overall deviation score (0–100), and show a confidence indicator using OCR quality + extraction coverage + model certainty + retrieval support.

Outputs : 1-page Document Summary + Template-Matching Report (deviations, severity, evidence spans, and remediation tips).

3.2 AGENT DESIGN

Each deed agent has three layers:

Template Layer (deterministic) : Match presence/absence of mandatory clauses against deed-type templates (e.g., attestation/witness, acceptance for gifts, chain of title, executor appointment, loan/security terms, discharge/release).

Consistency Layer : Verify internal consistency: name/date/property matching across clauses; mismatch and duplication detection; currency/amount sanity checks.

SLM Reasoning Layer (focused prompts) : Handle ambiguous or unusual language by comparing against template expectations (e.g., overbroad powers, hidden conditions, vague boundaries, unusual reservations).Prompts are short and deed-specific to keep costs low and reduce hallucinations.

Template Deviation Scoring & Confidence

Per-deviation score: Minor = 1, Moderate = 3, Major = 5 (weights adjustable per deed type).

Overall deviation score (0–100): normalize the weighted sum of deviations by document length and clause count.

Confidence (0,1): combine OCR quality × extraction coverage × model probability × retrieval support (if a law/guide note was retrieved).

Displayed as a numeric value plus a short label (e.g., “low confidence: poor scan on pages 3,4”).

Evaluation plan

Classification: accuracy, macro-F1 by deed type and by language (Si/Ta/En).

Extraction: token/character-F1 on fields.

Template matching: precision/recall/F1 for detecting deviations; span-F1 for cited evidence.

Retrieval (if enabled): top-k recall and MRR for law/guide passages.

Calibration: Brier score / Expected Calibration Error (ECE).

Runtime: average time per deed on CPU; target <2 minutes.

User study: reviewer time saved and trust score with/without evidence links.

Ablations: template-only vs SLM-only vs hybrid; effect of agent routing; OCR noise stress test.

3.3 DATA COLLECTION

Sources & ethics

Partners: notaries/conveyancers who can share de-identified sample deeds.

Public records: where lawful and permitted.

Synthetic: templated deeds with realistic variations to balance classes.

Privacy: remove names/IDs/addresses (replace with tags like <PERSON_A>), blur signatures, hash file names, and store locally.

Annotation

Team: two legal annotators + one adjudicators.

Labels: deed type; fields (parties, dates, property, amounts, witnesses, notary, encumbrances); template deviations (missing, extra, altered clauses); evidence spans (page + start–end character); language (Si/Ta/En mix); scan quality.

Tool: Label Studio

Agreement: target Cohen's $\kappa \geq 0.70$ on deviation categories and span overlap ≥ 0.6.

Split: stratified 70/15/15 (train/Val/test) by deed type and language.

Augment: add controlled scan noise (blur, skew), small paraphrases, and mixed-script tokens to match real documents.

3.4 SOFTWARE SOLUTION

Stack & deployment

Backend: Python (Fast API).

Models: small transformer encoders for classification/NER (e.g., legal-domain mini models) + a compact generative SLM for explanations; INT8/FP16 quantization with ONNX Runtime.

OCR: offline OCR with Sinhala/Tamil/English; preprocessing via OpenCV (binarize, denoise).

Vector store (optional RAG): FAISS or Qdrant (on-prem) for Sri Lankan statutes/circulars/practice notes.

Storage: Postgres (metadata), encrypted file store (MinIO/S3-compatible).

Frontend: simple Web UI (Stream lit or React) to upload files and view the report.

Orchestration: lightweight agent router (no heavy framework), async queues for batch jobs.

Security: runs on-prem, audit logs for every decision, role-based access, encryption at rest/in transit.

Packaging: Docker Compose for CPU-first; optional GPU.

Key modules

Validator: file size/type, page count, corruption check.

Parser: PDF to text + layout blocks; page map to keep page/line coordinates.

Classifier: deed type prediction with confidence and abstain option.

Extractor: regex + NER hybrid; cross-field checks (name/date/amount/property).

Deed agents (5): template matcher + SLM prompts; deviation emitter with evidence spans.

Scorer: severity weighting, overall deviation score, confidence calculator.

Reporter: Document Summary + Template-Matching Report (downloadable PDF/CSV/JSON).

Monitoring: runtime, failure reasons (e.g., “low OCR on p.4”), model drift alerts.

Success criteria

≥ 90% deed-type accuracy (macro-F1).

≥ 0.80 F1 on key fields (names, dates, property ID).

≥ 0.75 F1 on deviation detection; span-F1 ≥ 0.65 for evidence.

Review time reduced by ≥ 30% in pilot offices.

All processing offline/on-prem, with auditable outputs.

4.PROJECT REQUIREMENTS

4.1 FUNCTIONAL REQUIREMENTS

User& roles

Secure sign-in; roles: Reviewer (Notary/Conveyancer), Admin, Auditor.

Role-based access to uploads, results, and audit logs.

Document intake

Upload PDF/TIFF/JPG (single or batch).

Validate size, page count, corruption, and format; show clear errors.

Assign a unique Document ID and store metadata.

OCR & preprocessing

Detect scanned pages and run OCR for Sinhala/Tamil/English.

Preprocess: deskew, denoise, split pages, remove blank pages.

Save OCR quality score per page.

Deed classification

Predict one of five types: Power of Attorney, Transfer/Conveyance (Sale), Gift, Testamentary Deed (Last Will & Testament), Mortgage.

Show classification confidence and allow manual override.

Field extraction

Extract: parties, addresses, dates, property identifiers/plan numbers, consideration/amounts, encumbrances, witnesses, notary details, registration references.

Template matching (agent)

Route to the correct deed agent. Inside the agent, align clauses and extracted fields against deed-type templates (required, optional, prohibited).Emit deviations with severity (Minor/Moderate/Major), evidence span (page + line/text), and a short correction/remedy tip.

Scoring & confidence

Compute an overall deviation score (0–100) from weighted deviations.

Show a confidence indicator based on OCR quality, extraction coverage, model certainty, and (when enabled) retrieval support.

Retrieval grounding (optional)

Retrieve relevant Sri Lankan statutes/circulars/practice notes and cite them in findings to validate template rules.

Review workflow

Reviewer can accept/override deviations, add comments, and mark “resolved. “Changes are logged; system can learn from feedback (optional active-learning queue).

Reporting & export

Produce a Document Summary and a Template-Matching Deviation Report.

Export PDF/CSV/JSON; include audit trail and timestamps.

Operations

Job queue with status (Pending/Running/Done/Failed).

Email/in-app notifications on completion and errors.

API endpoint for batch ingestion (optional).

4.2 NON-FUNCTIONAL REQUIREMENTS

Accuracy targets

Deed-type macro-F1 ≥ 0.90.

Key fields (names, dates, property IDs) F1 ≥ 0.80.

Template deviation detection macro-F1 ≥ 0.75; evidence span F1 ≥ 0.65.

Performance

Latency < 2 minutes per 10-page deed on CPU-only workstation.

Batch throughput: ≥ 100 deeds/day on one node.

Security & privacy

On-prem/offline by default; no third-party data sharing.

Reliability

Job resume after crash; idempotent processing; checksum-based deduplication.

Uptime goal 99% in pilot.

Explainability

Every deviation must show rule/template check used and page/line evidence.

Usability

Clear deviation list with filters; one-click jump to evidence snippet; manual override.

Accessible UI; English first, Sinhala/Tamil UI labels in roadmap.

Maintainability

Modular micro-services; configuration by YAML; unit/integration tests ≥ 80% coverage; model and dataset versioning.

Portability

Docker zed; Linux/Windows servers; CPU first, GPU optional (CUDA if present).

Compliance & ethics

Human-in-the-loop; disclaimer: “decision support, not legal advice.”

4.3 EXPECTED TEST CASES

A. Intake & OCR

Upload valid PDF → passes validation; Document ID created.

Oversized file → clear error message.

Corrupted PDF → flagged at validation.

Low-quality scan → OCR runs; quality score < threshold; system shows "low confidence" banner.

Mixed Si/Ta/En page → correct language tokens extracted.

B. Classification

Transfer deed correctly classified (confidence ≥ threshold).

Ambiguous deed → model abstains; asks for manual confirmation.

Manual override updates downstream agent and report.

C. Field Extraction & Consistency

Parties/dates/amounts/witnesses detected with page references.

Name mismatch across clauses → consistency deviation flagged.

Invalid date order (execution after registration) → template violation raised.

Missing notary details → required field missing deviation.

D. Agent Template Checks (per deed)

Gift deed without acceptance wording → critical deviation flagged with evidence.

Power of Attorney with over-broad scope → non-standard clause flagged by SLM + snippet.

Mortgage missing release/discharge terms → major deviation.

Transfer deed lacking chain-of-title continuity → template gap with list of missing links.

Testamentary deed without executor appointment → required clause missing deviation.

E. Retrieval & Grounding

When RAG enabled, retrieved statute confirms template expectation; citation shown.

If retrieval fails, deviation still emitted but marked “no external authority found.”

F. Scoring, Confidence, and Reporting

Multiple deviations aggregate to overall deviation score (0–100).

Evidence links open exact page/line.

Report exports (PDF/CSV/JSON) match UI contents.

Reviewer edits recorded in audit log with user/time.

G. Robustness & Security

Large 200-page deed completes within SLA.

Concurrent uploads (≥10) processed without data mixing.

Unauthorized user cannot view/download reports.

System resumes job after intentional crash.

4.4 FEASIBILITY STUDY

Technical feasibility

The solution uses proven building blocks: offline OCR (Sinhala/Tamil/English), small transformer models (quantized INT8/FP16) for classification/NER, and a compact generative SLM for template deviation explanations. A lightweight rule engine plus focused prompts keeps compute low and improves explainability. Retrieval uses an on-prem vector store (FAISS/Qdrant). All services run in Docker on a mid-range CPU server; GPU is optional. Main challenge: mixed-script OCR and noisy scans. Mitigation: domain-specific OCR tuning, post-OCR cleanup, and evidence-span verification.

Operational feasibility

Notaries and conveyancers can adopt the tool because outputs show page/line evidence of template deviations and allow manual override. Training needs are modest (2–3 hours). A feedback screen lets reviewers correct extracted fields and flagged deviations, which we can use to refine models and rules over time.

Economic feasibility

Small models and on-prem deployment avoid high cloud costs. A single workstation can process 100 deeds/day, cutting review time and cost. Main expenses are annotation effort and initial setup; these are one-time or periodic.

Legal & privacy feasibility

All processing is local; data is encrypted and access controlled. The system keeps an audit trail and displays a clear disclaimer (decision support, not legal advice). De-identification is applied for any training data.

Schedule feasibility (indicative 8,9 months)

M1,M2: Template library & deviation taxonomy, intake/OCR module, annotation protocol.

M3,M4: Classifier + field extractor MVP; start dataset labeling.

M5: Deed agents (rules + prompts) and deviation scoring; reporting UI.

M6: Confidence & evidence spans; optional RAG integration.

M7: End-to-end tests; robustness & security; pilot in one office.

M8,M9: Fixes, documentation, and handover.

Key risks & mitigations

Data scarcity: use de-identified samples + synthetic template variants; active learning to prioritize labeling.

OCR errors: quality scoring + human review of low-confidence pages.

Template drift: version templates; schedule quarterly updates with a legal advisor.

User trust: always show evidence spans and confidence; keep human-in-the-loop.

5.SYSTEM DIAGRAM AND GANTT CHART

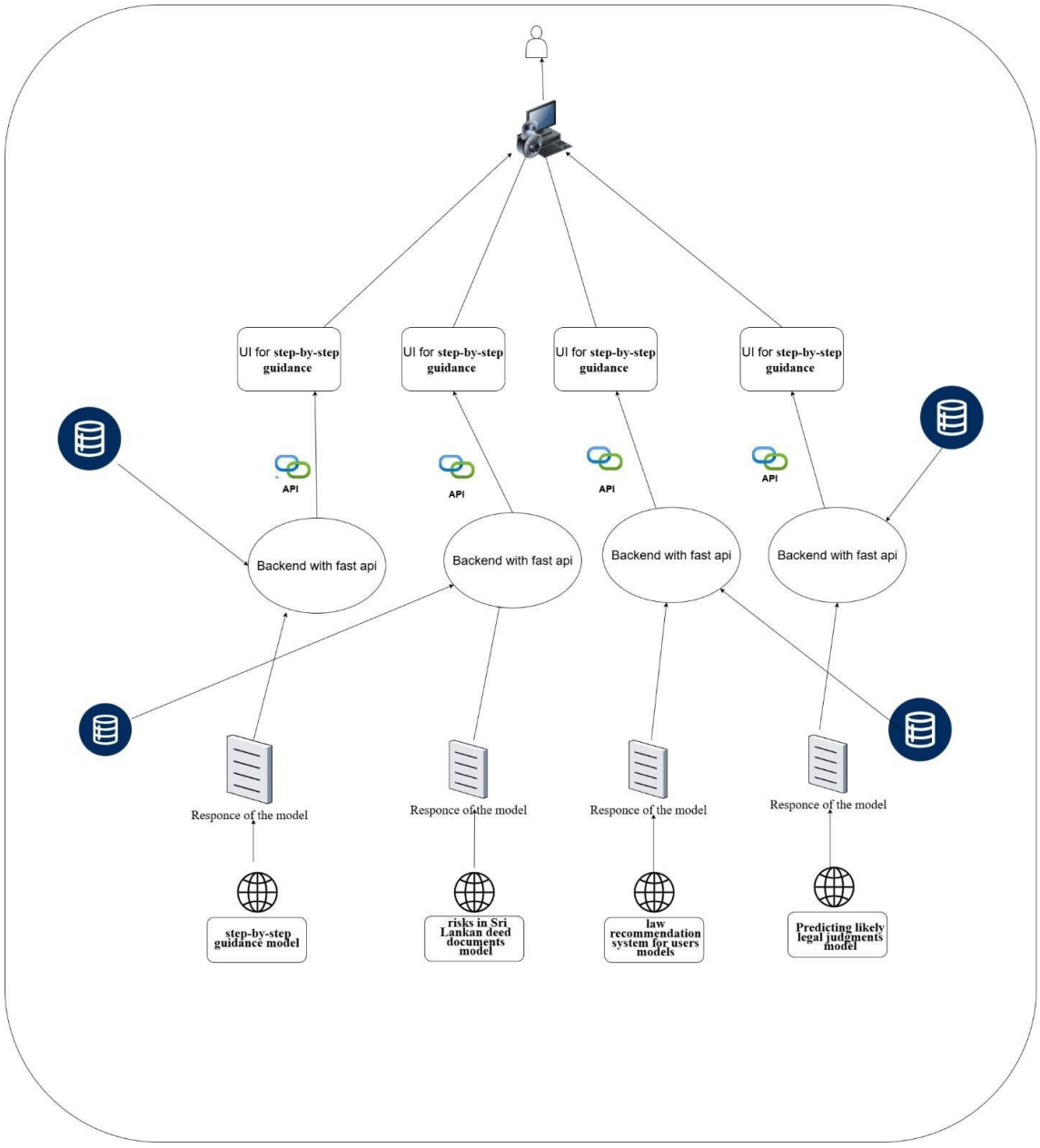


Figure 1:System Diagram for Relevance and Abuse Detection

5.1 TECHNICAL DAIGRAM

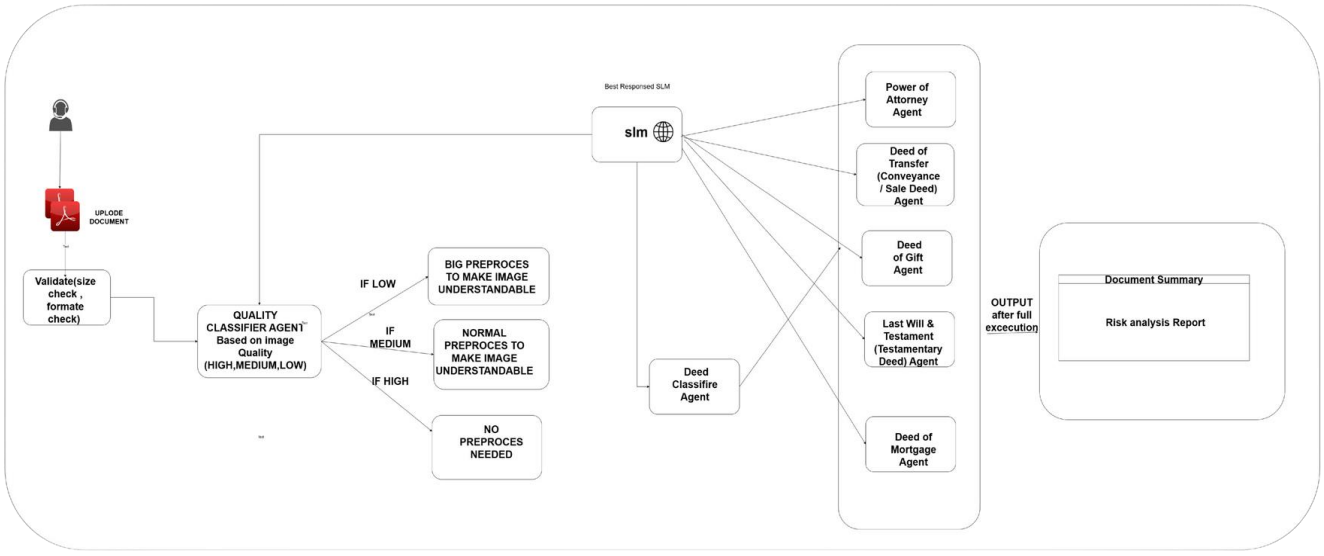


Figure 2:Technical diagram

The diagram depicts the complete workflow of the deed evaluation system. Users upload a legal document, which is validated for format and size. The Deed Classifier Agent and SLM collaboratively determine the deed type. The system then produces a Document Summary and a Risk Analysis Report, highlighting potential legal risks with references to applicable laws. The design integrates AI and rule-based verification for an efficient, reliable, and user-friendly document assessment process.

5.2 GANNT CHART

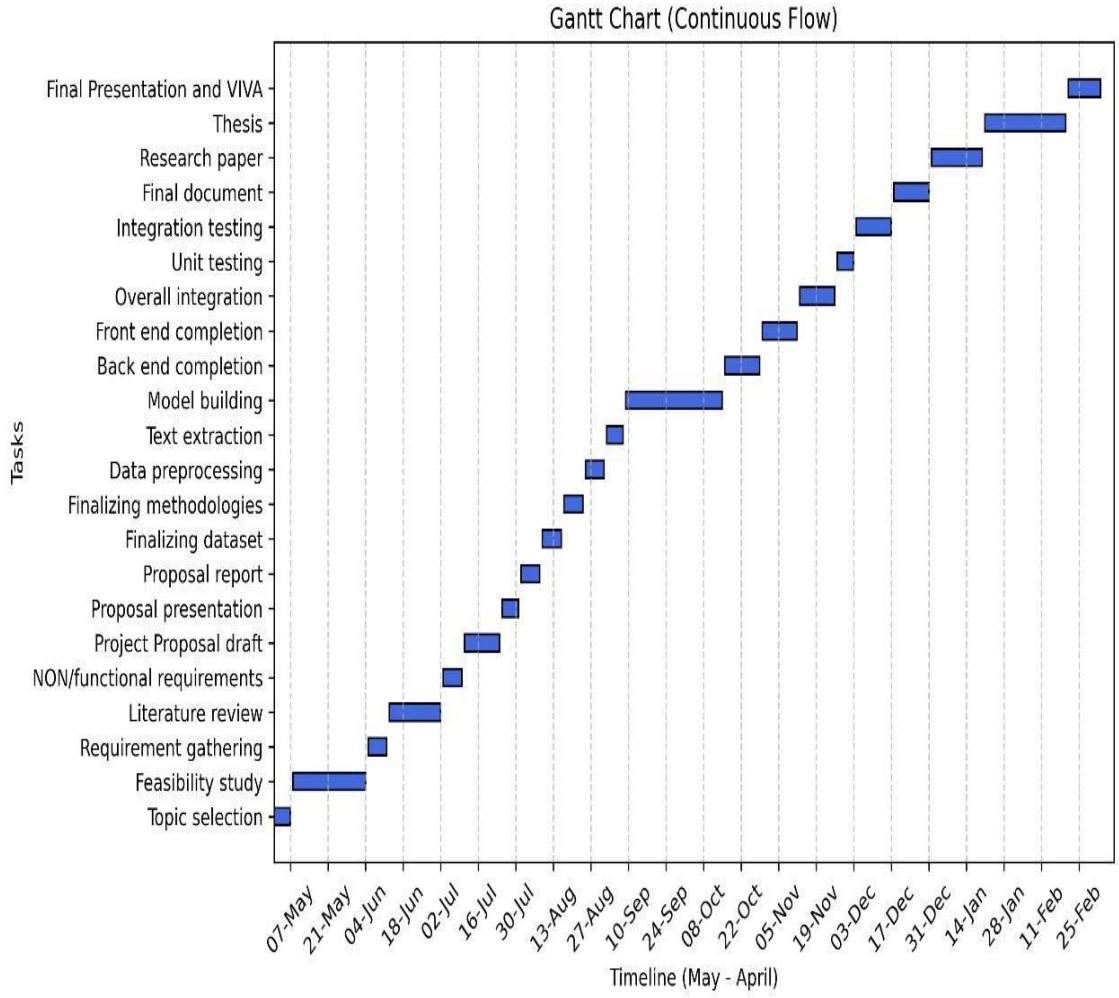


Figure 3:Gannt chart

6.BUDGET AND BUDGET JUSTIFICATION

The Smart Deed Document Template Matching budget is provided in LKR and includes key expenditures to support sustainable implementation within Sri Lanka's

court framework. Major allocations are as follows:

Category	Cost (LKR)	Justification
Web hosting & deployment infrastructure	5,000	to provide secure, on-premises or court-network access for authorized users.
Development Tool Cost	4,000	Includes licenses for programming tools such as Google colab, VScode
Internet Cost	6,500	to enable dataset updates, legal reference retrieval (when permitted), and remote collaboration during the pilot.
Cloud Services (GCP)	15,000	Estimated GCP consumption (e.g., e2-micro instances @ ~LKR 2.4/hour, 500 hours/month for 9 months) and storage requirements.
Stationary Cost	5,000	Covers paper, printing, physical document scanning for data extraction and office supplies for documentation and meetings.
Miscellaneous	3,000	Allows for unexpected costs such as travel and small equipment.
Total	**38,500**	

Table 5:Budget And Budget Justification

6.1 COMMERCIALIZATION

Market Opportunity

The legal sector in Sri Lanka, like in many developing countries, faces challenges with deed verification, contract management, and compliance checking. Errors in documentation and lack of affordable expert review often lead to disputes, delays, and

additional costs. At the same time, the adoption of digital tools in legal practice is growing steadily. This creates a strong opportunity for an AI-powered deed template matching system that can provide quick, affordable, and reliable insights by detecting deviations from standard legal deed structures.

The primary target market includes law firms, notaries, banks, real estate companies, and government agencies involved in property transactions. In addition, individual users who wish to validate deeds against standard legal formats before signing contracts represent a large untapped segment. With the increasing digitalization of legal services and rising trust in AI-assisted platforms, there is significant room for commercialization both locally and regionally.

Revenue Model

1. Subscription Plans – Law firms, notaries, and organizations can subscribe on a monthly or yearly basis for unlimited document validation.
2. Pay-Per-Use Model – Individual users or small firms can pay per document for one-time validation without long-term commitment.
3. Enterprise Licensing – Large institutions such as banks and government bodies can purchase enterprise licenses with additional customization, integration, and support services.
4. Freemium Model – A free version with basic checks (e.g., field extraction and simple summaries) can attract users, while advanced template matching and detailed deviation reports are offered in a premium package.

7.CONCLUSION

This research shows that small language models (SLMs) combined with rule-based template matching can be applied in a practical and trustworthy way for the Sri Lankan legal domain. We designed a full pipeline that takes a deed, checks document quality, extracts key fields, routes it to the correct deed agent, and then runs a template compliance algorithm inside that agent. The algorithm mixes clear rule checks (must-have clauses

and cross-checks) with focused SLM reasoning for ambiguous or unusual wording. The system produces two outputs that reviewers can use immediately: a Document Summary and a Template Deviation Report with page/line evidence and a confidence indicator. Pilots with both clean PDFs and noisy scans suggest that the agentic design is effective. Having a separate agent for each deed type lets the system use narrow prompts and deed-specific checklists. The tool consistently flagged high-impact deviations such as missing attestation details, name mismatches, unclear property boundaries, absent acceptance wording in gifts, over-broad powers in powers of attorney, and incomplete release terms in mortgages. Reviewers reported that evidence links made verification faster and easier, and the confidence label helped them judge when a page needed manual attention. Compared with large cloud models, our approach is lighter, cheaper, and privacy friendly. It runs on-prem, supports Sinhala/Tamil/English through OCR and extraction, and is explainable: every alert shows the text span and the rule or prompt that triggered it. The method also fits with our wider program: SLMs for Property/Family Law, labor-law recommendation services, and outcome prediction from past cases creating a consistent platform for legal AI in Sri Lanka. There are limits. High-quality, Sri Lanka-specific deed datasets are scarce; scanned pages can be noisy; and laws and practice notes change over time. Mixed scripts (Si/Ta/En) still challenge OCR and extraction. Fairness must be checked across languages, regions, and older deed formats.

Next steps include:

1. growing a de-identified Sri Lankan deed dataset with template labels and evidence spans,
2. fine-tuning OCR and extraction for Sinhala/Tamil and stamps/seals,

3. adding time-aware retrieval grounding to align checks with updated statutes and circulars,
4. capturing reviewer feedback for active learning,
5. integrating optional registry lookups (encumbrances, caveats), and
6. publishing an evaluation protocol with span-level metrics and calibration tests.

In short, a rule-guided SLM with deed-specific template agents is a practical path to faster, more reliable deed review in Sri Lanka, delivering explainable results that legal professionals can trust.

REFERENCES

[1] D. Chakrabarti et al., "Use of Artificial Intelligence to Analyze Risk in Legal Documents for a Better Decision Support," arXiv.org, 2019. https://arxiv.org/abs/1912.01111

[2] "Contract Eval: Benchmarking LLMs for Clause-Level Legal Risk Identification in Commercial Contracts," Arxiv.org, 2025. https://arxiv.org/html/2508.03080 (accessed Aug. 28, 2025).

[3] "A Comprehensive Framework for Reliable Legal AI: Combining Specialized Expert Systems and Adaptive Refinement," Arxiv.org, 2023. https://arxiv.org/html/2412.20468v1 (accessed Aug. 28, 2025).

[4] R. C. Barron, M. E. Eren, O. M. Serafimova, C. Matuszek, and B. S. Alexandrov, "Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization," arXiv.org, 2025. https://arxiv.org/abs/2502.20364

[5] L. Wan, G. Papageorgiou, M. Seddon, and M. Bernardoni, "Long-length Legal Document Classification," arXiv.org, 2019. https://arxiv.org/abs/1912.06905

[6] F. Wei, H. Qin, S. Ye, and H. Zhao, "Empirical Study of Deep Learning for Text Classification in Legal Document Review," 2018 IEEE International Conference on Big Data (Big Data), Dec. 2018, doi: https://doi.org/10.1109/bigdata.2018.8622157.

[7] F. Ariai and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," arXiv.org, 2024. https://arxiv.org/abs/2410.21306

[8] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, and M. J. Bommarito II, "Natural Language Processing in the Legal Domain," arXiv.org, Feb. 23, 2023. https://arxiv.org/abs/2302.12039

[9] M. Akter, E. Çano, E. Weber, D. Dobler, and I. Habernal, “A Comprehensive Survey on Legal Summarization: Challenges and Future Directions,” arXiv.org, 2025. https://arxiv.org/abs/2501.17830

[10] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, “Legal Judgment Prediction via Topological Learning,” ACLWeb, Oct. 01, 2018. https://aclanthology.org/D18-1390/

[11] W. Yang, W. Jia, X. Zhou, and Y. Luo, “Legal Judgment Prediction via MultiPerspective Bi-Feedback Network,” arXiv (Cornell University), pp. 4085–4091, Jul. 2019, doi: https://doi.org/10.24963/ijcai.2019/567.

[12] P. Madambakam, S. Rajmohan, H. Sharma, and Gupta, “SLJP: Semantic Extraction based Legal Judgment Prediction,” arXiv.org, 2023. https://arxiv.org/abs/2312.07979 (accessed Aug. 29, 2025).

[13] S. Wong, C. Zheng, X. Su, and Y. Tang, “Construction contract risk identification based on knowledge-augmented language model,” arXiv.org, 2023. https://arxiv.org/abs/2309.12626

[14] N. Limsopatham, “Effectively Leveraging BERT for Legal Document Classification,” ACLWeb, Nov. 01, 2021. https://aclanthology.org/2021.nllp1.22/#:~:text=Bidirectional%20Encoder%20Re presentations%20from%20Transf ormers%20%28BERT%29%20has%20achieved

[15] “ACORD: An Expert-Annotated Dataset for Legal Contract Clause Retrieval,” Arxiv.org, 2023. https://arxiv.org/html/2501.06582v2 (accessed Aug. 29, 2025).

[16] "ACORD: An Expert-Annotated Dataset for Legal Contract Clause Retrieval," Arxiv.org, 2023. https://arxiv.org/html/2501.06582v2 (accessed Aug. 29, 2025).

[17] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," arXiv:2010.02559 [cs], Oct. 2020, Available: https://arxiv.org/abs/2010.02559

[18] S. Comu, A. Y. Elibol, and B. Yucel, "A risk assessment model of commercial real estate development projects in developing countries," Journal of Construction Engineering, Management & Innovation, vol. 4, no. 1, pp. 52–67, Mar. 2021, doi: https://doi.org/10.31462/jcemi.2021.01052067.

[19] "Explainable Artificial Intelligence Credit Risk Assessment using Machine Learning," Arxiv.org, 2020. https://arxiv.org/html/2506.19383v1

[20] P. Belcak et al., "Small Language Models are the Future of Agentic AI," arXiv.org, 2025. https://arxiv.org/abs/2506.02153

[21] "LawGPT: A Chinese Legal Knowledge-Enhanced Large Language Model," Arxiv.org, 2023. https://arxiv.org/html/2406.04614v1 (accessed Aug. 19, 2025).

[22] R. C. Barron, M. E. Eren, O. M. Serafimova, C. Matuszek, and B. S. Alexandrov, "Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization," arXiv.org, 2025. https://arxiv.org/abs/2502.20364

[23] K. Stødle, R. Flage, S. Guikema, and T. Aven, “Artificial intelligence for risk analysis—A risk characterization perspective on advances, opportunities, and limitations,” Risk analysis, Apr. 2024, doi: https://doi.org/10.1111/risa.14307.

[24] “An Artificial Intelligence based Analysis in Legal domain,” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 2S2, pp. 1046–1051, Dec. 2019, doi: https://doi.org/10.35940/ijitee.b1113.1292s219.