UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE
KATHMANDU

# 6CS007

# PROJECT AND PROFESSIONALISM

## PROJECT PROPOSAL REPORT-

## NYAYAMITRA: RAG based AI for Nepali legal Q&A

| | |
|---|---|
| University Id | : 2408239 |
| Group | : L6CG15 |
| Reader | : Mr. Akash Adhikari |
| Supervisor | : Mr. Yamu Poudel |
| Student Name | : Nikisha Shrestha |
| Course | : BSc (Hons) Computer Science |
| Submitted on | : 09/15/2025 |
| Word Count | : 2433 |

## Acknowledgement

I would like to sincerely thank my supervisor, Mr. Yamu Poudel, and my reader, Mr. Akash Adhikari, for their valuable suggestions and guidance during the preparation of this proposal. Their feedback has been very helpful in shaping and improving this work.

# Contents

# Table of Figures

# Table of Tables

**Project Title -** NyayaMitra: RAG based AI for Nepali legal Q&A

## 1. Problem Statement

Lawyers and legal professionals are often burdened by the need to manually analyze large volumes of lengthy and complex documents such as case laws, contracts, statutes and regulations, a process that is both time-intensive and inefficient. Existing AI systems like ChatGPT and Claude, while powerful in general domains, lack the legal domain training required to deliver reliable outputs for specialized use cases. This research therefore seeks to address the gap by developing an AI-powered legal chatbot based on LEGAL-BERT, enabling accurate and efficient document analysis that aligns with the increasing demand for digital transformation in the legal sector.

## 2. Academic Questions

**R01**: How can AI and NLP models leverage domain-specific techniques and embeddings to develop a legal chatbot that answers specific questions as well as cite relevant information from PDF documents?

**R02**: How can large legal documents be automatically summarized while preserving key information and context for legal professionals?

**R03**: How can semantic similarity and cross-PDF analysis be used within a legal chatbot to improve comparative legal research across multiple documents?

## 3. Aims

To develop an AI-powered legal chatbot that enables legal professionals to upload PDF documents, ask specific questions, receive citation-first answers, and obtain accurate summaries of the documents.

## 4. Objectives

1. To design a user-friendly platform that allows legal professionals to upload, manage and process multiple PDF documents efficiently.
2. To implement a domain-specific NLP model for accurate legal question answering and entity recognition.
3. To develop a document summarization module that generates concise and context-preserving summaries of long legal texts.
4. To integrate semantic similarity and cross-PDF analysis for comparative legal research and case law referencing
5. To ensure data privacy, security, and cloud accessibility through end-to-end encryption and SaaS deployment.

## 5. Background of Study

The integration of artificial intelligence (AI) and natural language processing (NLP) into document analysis has gained significant momentum in recent years. Various systems have been developed to enable chatbot-assisted query answering and document summarization from PDF-based documents across domains such as automotive manuals, human resources, education and finance (Medeiros, et al., 2023) (Shah, Ryali, & Venkatesh, 2024). Models like BERT and GPT-3 have demonstrated strong capabilities in semantic understanding and question answering, while frameworks such as Langchain and LlamaIndex simplify document processing and interaction with textual content (Medeiros, et al., 2023) (Patil, et al., 2024).

Similar efforts have been made to adapt these models for tasks specific to legal practice. General-purpose models like BERT and GPT-3 often fail to provide precise answers to complex legal queries. To address this, researchers have introduced legal-domain pretrained models, such as LEGAL-BERT (Silno, Falco, Croce, & Rosso, 2025), which are fine-tuned on legal corpora in English. However, this creates a limitation in regions like Nepal, where most legal documents are written in Nepali, a language not supported by LEGAL-BERT. To overcome this gap, NepKanun, a retrieval-augmented generation (RAG) based legal assistant, was developed using LLaMA 3.2 3B, fine-tuned on Nepali legal datasets (NepKanun: A RAG-Based Nepali Legal Assistant, 2025).

While NepKanun demonstrates promising results for Nepali legal texts, it lacks the ability to process English legal documents effectively due to its language-specific training (NepKanun: A RAG-Based Nepali Legal Assistant, 2025). This highlights the need for a hybrid bilingual approach, where English legal documents are handled by LEGAL-BERT while Nepali legal texts are processed using a fine-tuned LLaMA model. Such a system would ensure robust bilingual support, accurate legal information retrieval, and efficient summarization, addressing a critical gap in existing research and offering practical benefits to Nepali legal professionals.

## 6. Feature Breakdown

### 1. Multiple Document Handling

- Upload, parse, and manage multiple PDFs simultaneously for efficient processing

### 2. Bilingual Support

- Supports Nepali and English legal documents seamlessly

### 3. Legal Question Answering (Q & A)

- Provides accurate, citation-first answers using advanced legal NLP models

### 4. Document Summarization

- Generated concise, coherent summaries of long legal documents at both chunk and document level

### 5. Named Entity Recognition (NER)

- Identifies and classifies entities (e.g., name of people, organizations, locations) within legal documents to ensure precise citations

### 6. Similarity Estimation & Cross-Referencing

- Measures semantic similarity between legal documents or cases which is essential for case law comparison
- Enables cross-PDF referencing and case law comparison

### 7. Cloud Deployment (SaaS)

- Fully web-accessible platform hosted on the cloud for scalability and ease of use

### 8. Data privacy & Security

- End-to-end encryption ensures sensitive legal documents remain secure throughout upload, processing, and storage

## 7. FDD (Functional Decomposition Diagram) & Functionality

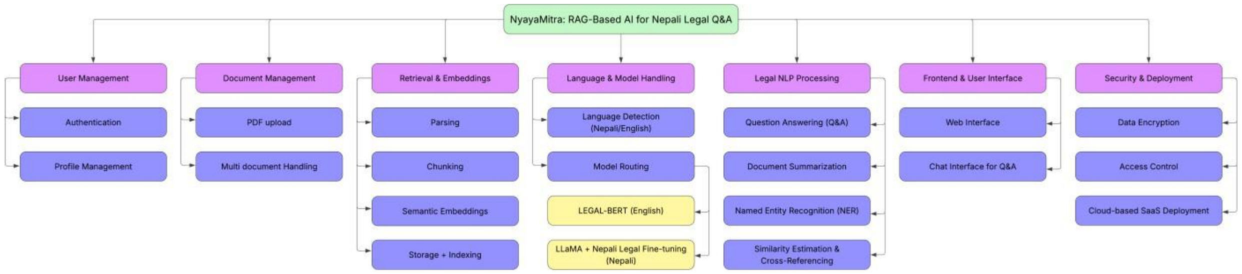


*Figure 1: Functional Decomposition Diagram (FDD)*

### 1. User Management Module

- Authentication: Handles login and signup process.
- Profile Management: Allows users to manage their account and track previously uploaded documents.

### 2. Document Management Module

- PDF Upload: User can upload multiple PDFs.
- Multi-Document Handling: Allows simultaneous processing of multiple PDFs, supporting batch queries and cross-document comparisons.

### 3. Retrieval & Embeddings Module

- Parsing: The system automatically parses the uploaded PDF documents.
- Chunking: Divides documents into structured chunks (e.g., sections, articles) for efficient retrieval.
- Semantic Embeddings: Creates embeddings for document chunks for semantic similarity calculations.
- Storage + Indexing: Stores documents in the database and indexes them for fast retrieval.

### 4. Language & Model Handling

- Language Detection: Automatically detects whether a document is Nepali or English.
- Model Routing: Routes English documents to LEGAL-BERT and Nepali documents to LLaMA fine-tuned on Nepali legal texts.

5. **Legal NLP Processing Module**

- Question Answering (Q&A): Provides citation-first answers to user queries using domain-specific models.
- Document Summarization: Generates coherent summaries at both chunk and document levels.
- Named Entity Recognition (NER): Identifies and classify legal entities such as case names, statutes, sections, organizations.
- Similarity Estimation & Cross-Referencing: Measures semantic similarity across legal documents and supports comparative research.

6. **Frontend & User Interface Module**

- Web Interface: Provides a web-accessible interface for uploading documents and chatting with the assistant.
- Chat Interface for Q&A: Allows real-time question answering with citation-first responses.

7. **Security & Deployment Module**

- Data Encryption: Ensures end-to-end encryption for uploaded legal documents.
- Access Control: Manages permissions and prevents unauthorized access.
- Cloud-based SaaS Deployment: Deploys the system on the cloud for scalability and cross-device access.

## 8. Software Development Life Cycle (SDLC)

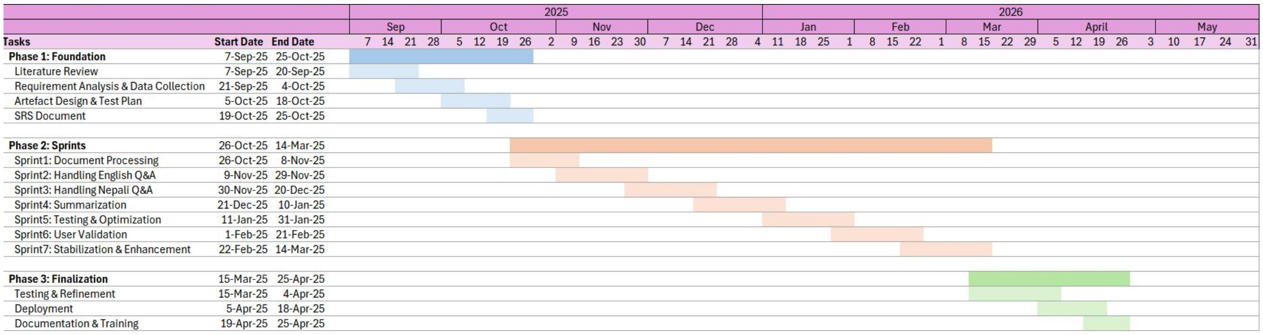
The Agile methodology will be followed for the development of NyayaMitra. Iterative development, continuous feedback, and incremental improvements will be emphasized, allowing the system to be adapted based on evolving user requirements. Features such as PDF parsing, Q&A, and summarization will be refined through regular feedback, ensuring a user-friendly and reliable product is delivered in shorter cycles.

## 9. Tech Stack

| Category | Component | Tools | Reason |
|---|---|---|---|
| AI & NLP | Q&A + NER | LEGAL-BERT | -pre-trained on English legal text, providing accurate Q&A and NER |
| | | LLaMA 3.2 3B model fine-tuned on Nepali legal documents | -allows Nepali legal document understanding |
| | Embeddings & Similarity Detection | Sentence-BERT (multilingual variant) | -creates semantic embeddings for PDF chunks, enabling cross-PDF retrieval and similarity detection, and supports Nepali and English |
| | Pipeline/PDF Handling | Langchain + LlamaIndex | -automates PDF parsing, chunking, indexing, and retrieval, integrating all models seamlessly |
| Backend & Frontend | Backend | FastAPI | -lightweight, high-performance framework for building APIs<br>-integrates well with Python-based AI models |
| | Frontend | React | -enables creation of dynamic and interactive user interfaces<br>-component-based architecture allows modular and maintainable code<br>-large ecosystem simplifies integration with APIs and AI models |
| Database & Deployment | Database | ChromaDB (Vector DB) | -stores embedding efficiently for semantic search |
| | | PostgreSQL (RelationalDB) | -manages structured data like user credentials and profiles |
| | Deployment | Cloud-based SaaS with end-to-end encryption (AWS) | -provides scalable, secure, and accessible service<br>-AWS supports GPU hosting for AI models and managed relational databases. |

*Table 1: Tech Stack*

## 10. Gantt Chart



| Tasks | Start Date | End Date |
|---|---|---|
| **Phase 1: Foundation** | 7-Sep-25 | 25-Oct-25 |
| Literature Review | 7-Sep-25 | 20-Sep-25 |
| Requirement Analysis & Data Collection | 21-Sep-25 | 4-Oct-25 |
| Artefact Design & Test Plan | 5-Oct-25 | 18-Oct-25 |
| SRS Document | 19-Oct-25 | 25-Oct-25 |
| | | |
| **Phase 2: Sprints** | 26-Oct-25 | 14-Mar-25 |
| Sprint1: Document Processing | 26-Oct-25 | 8-Nov-25 |
| Sprint2: Handling English Q&A | 9-Nov-25 | 29-Nov-25 |
| Sprint3: Handling Nepali Q&A | 30-Nov-25 | 20-Dec-25 |
| Sprint4: Summarization | 21-Dec-25 | 10-Jan-25 |
| Sprint5: Testing & Optimization | 11-Jan-25 | 31-Jan-25 |
| Sprint6: User Validation | 1-Feb-25 | 21-Feb-25 |
| Sprint7: Stabilization & Enhancement | 22-Feb-25 | 14-Mar-25 |
| | | |
| **Phase 3: Finalization** | 15-Mar-25 | 25-Apr-25 |
| Testing & Refinement | 15-Mar-25 | 4-Apr-25 |
| Deployment | 5-Apr-25 | 18-Apr-25 |
| Documentation & Training | 19-Apr-25 | 25-Apr-25 |

*Figure 2: Gantt Chart*

## 11. Testing Methodologies

1. **Unit Testing:** Unit Testing is conducted to test individual models such as PDF parsing, text embedding, AI query handling, and database operations. The purpose of this test is to verify that each component works correctly in isolation. Tools: pytest

2. **Integration Testing:** Integration testing is conducted to test the interaction between modules, e.g., how document parsing works with embedding storage and retrieval. The purpose of this test is to ensure that integrated components communicate and operate correctly together. Tools: pytest, Postman (for API endpoints)

3. **System Testing:** System testing is conducted to test the complete system including the frontend, backend, and AI processing pipeline. The purpose of this test is to verify the system as a whole meets functional and non-functional requirements. Tools: Selenium (for automated UI testing), JMeter (for performance testing)

4. **User Acceptance Testing (UAT):** Real end-users (lawyers, legal professionals) test the system by uploading PDFs, asking questions, and receiving summaries.

The purpose of this test is to ensure system meets user expectations and is user-friendly. The approach taken in this test is to collect feedback and make iterative improvements before final deployment.

## 12. Future Enhancements

1. **Advanced Summarization Models:** Incorporating more robust summarization techniques, such as the Longformer Encoder-Decoder (LED) or other transformer-based models, to handle lengthy and complex legal documents more effectively.

2. **Enhanced Reasoning Capabilities:** Integrating models with stronger general reasoning abilities, such as the GPT API, to improve contextual understanding and provide more accurate legal insights.

3. **Faster Information Retrieval:** Implementing efficient retrieval techniques like BM25 or hybrid dense-sparse retrieval methods to reduce query latency and improve response accuracy.

4. **Domain-Specific Knowledge Integration:** Expanding the system’s legal knowledge base through the use of legal synonym dictionaries and ontologies, enabling better interpretation of nuanced legal terms and concepts.

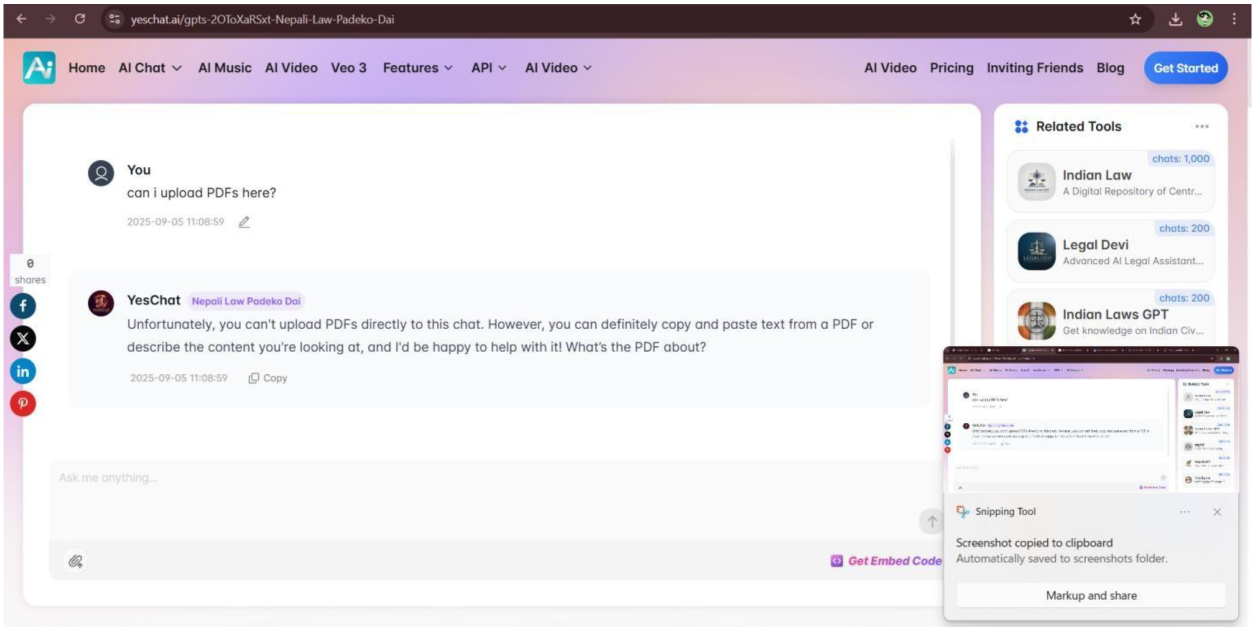
## 13. Similar Works

1. **Nepali Law Padheko Dai**

   ‘Nepali Law Padeko Dai’ is an AI-driven legal assistant designed to provide comprehensive access to Nepali legal knowledge. It supports advanced functionalities including legal information retrieval, taxation guidance, interpretation of statutory provisions, and constitutional rights analysis. The assistant is intended for legal professionals, academics, business stakeholders, and the general public.

The platform is accessible in two ways, each using a different underlying model. On YesChat.ai, ‘Nepali Law Padeko Dai’ runs on ChatGPT-4, providing users with document summarization and context-aware legal query answering (Bijukchhe). When accessed as a Custom GPT within ChatGPT (via the “Explore GPTs” or Search GPT feature), the assistant runs on GPT-5, leveraging improved reasoning and contextual understanding while maintaining the same core functionalities (Bijukchhe, Nepali Law Padheko Dai). A separate website isn’t needed for this application as users can interact with the assistant either through YesChat.ai or ChatGPT. Access via yeschat.ai or ChatGPT.

The assistant’s document processing pipeline combines multi-document ingestion and retrieval-augmented generation (RAG) techniques to extract relevant legal information efficiently. This architecture allows seamless integration of the assistant into broader workflows, enabling both interactive querying and structured legal document analysis.



You

can i upload PDFs here?

2025-09-05 11:08:59

YesChat Nepali Law Padeko Dai

Unfortunately, you can't upload PDFs directly to this chat. However, you can definitely copy and paste text from a PDF or describe the content you're looking at, and I'd be happy to help with it! What's the PDF about?

2025-09-05 11:08:59 Copy

Screenshot copied to clipboard
Automatically saved to screenshots folder.

*Figure 3: Accessing ‘Nepali Law Padheko Dai’ via yeschat.ai*

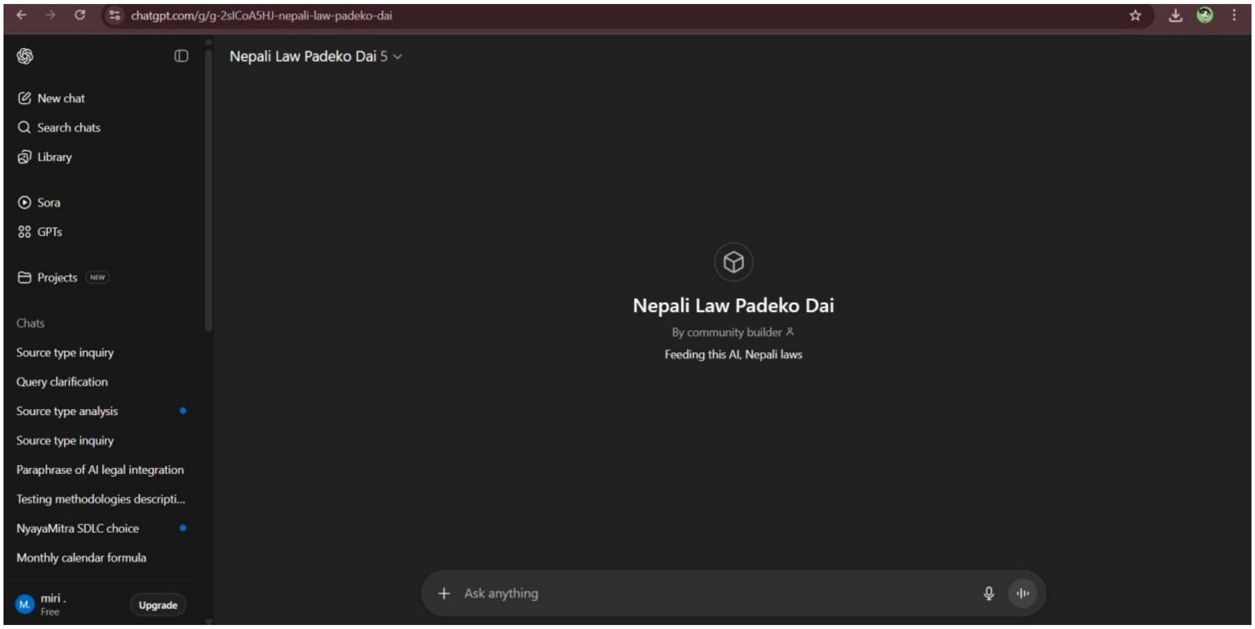
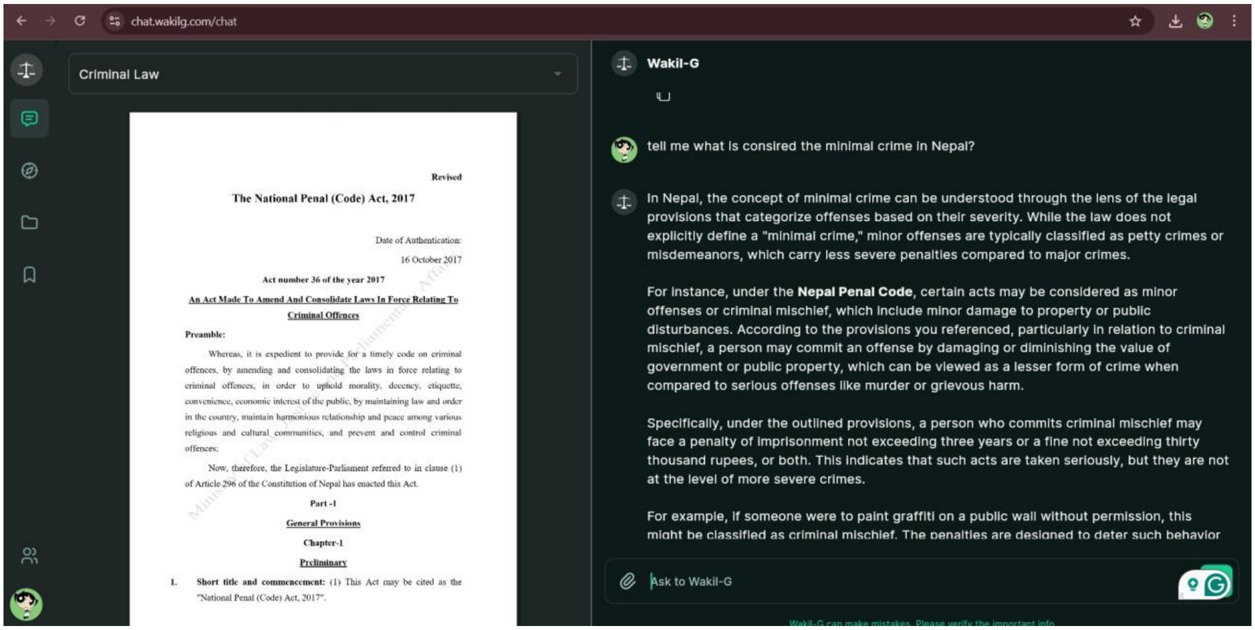


*Figure 4: Accessing 'Nepali Law Padheko Dai' via ChatGPT*

## 2. Wakil-G

Wakil-G is an AI-powered legal assistant that provides general legal information based on Nepali law, including the Constitution and official statutes. It offers 24/7 access to legal guidance or legal professionals, students, business owners, and the general public. Users interact with the assistant through a conversational interface, asking questions on topics such as constitutional rights, property law, family law, and business regulations. While Wakil-G delivers reliable legal information, it does not support document uploads or provide case-specific legal advice. Compared to Nepali Law Padeko Dai, Wakil-G focuses on conversational guidance rather than document analysis or multi-platform integration (Basnet, 2025). Access via Wakil-G


chat.wakilg.com/chat
Criminal Law
Wakil-G
Ask to Wakil-G
Wakil-G can make mistakes. Please verify the important info.

Revised

The National Penal (Code) Act, 2017

Date of Authentication:
16 October 2017

Act number 36 of the year 2017

An Act Made To Amend And Consolidate Laws In Force Relating To Criminal Offences

Preamble:

Whereas, it is expedient to provide for a timely code on criminal offences, by amending and consolidating the laws in force relating to criminal offences, in order to uphold morality, decency, etiquette, convenience, economic interest of the public, by maintaining law and order in the country, maintain harmonious relationship and peace among various religious and cultural communities, and prevent and control criminal offences;

Now, therefore, the Legislature-Parliament referred to in clause (1) of Article 296 of the Constitution of Nepal has enacted this Act.

Part -1

General Provisions

Chapter-1

Preliminary

1. Short title and commencement: (1) This Act may be cited as the "National Penal (Code) Act, 2017".

tell me what is consired the minimal crime in Nepal?

In Nepal, the concept of minimal crime can be understood through the lens of the legal provisions that categorize offenses based on their severity. While the law does not explicitly define a "minimal crime," minor offenses are typically classified as petty crimes or misdemeanors, which carry less severe penalties compared to major crimes.

For instance, under the **Nepal Penal Code**, certain acts may be considered as minor offenses or criminal mischief, which include minor damage to property or public disturbances. According to the provisions you referenced, particularly in relation to criminal mischief, a person may commit an offense by damaging or diminishing the value of government or public property, which can be viewed as a lesser form of crime when compared to serious offenses like murder or grievous harm.

Specifically, under the outlined provisions, a person who commits criminal mischief may face a penalty of imprisonment not exceeding three years or a fine not exceeding thirty thousand rupees, or both. This indicates that such acts are taken seriously, but they are not at the level of more severe crimes.

For example, if someone were to paint graffiti on a public wall without permission, this might be classified as criminal mischief. The penalties are designed to deter such behavior

*Figure 5: Wakil-G Interface*

## 14. Limitations

1. **Time Constraints:** Due to limited project duration, advanced fine-tuning of large-scale models and extensive evaluation across diverse legal datasets may not be fully achieved.
2. **Data Availability:** Access to comprehensive and well-annotated Nepali datasets is limited, which may affect the accuracy of domain-specific responses.

3. **Scope Restriction:** The system is primarily focused on legal documents Q&A and summarization; it does not cover advanced legal reasoning, predictive analytics, or full case law research at this stage.

4. **Resource Constraints:** Training and deploying larger models require high computational resources (GPUs/TPUs), which may not be fully available within the project setup.

## 15. Conclusion

This project aims to develop an AI-powered legal PDF assistant capable of answering queries and summarizing lengthy legal documents, with a particular focus on Nepali and English contexts. By combining natural language processing, retrieval mechanisms, and domain-specific enhancements, the system seeks to reduce the time and effort legal professionals spend reviewing documents. Although limited in scope and resources, the project demonstrates how AI can support the legal sector in Nepal and beyond, laying the foundation for future research and more sophisticated legal AI applications.

## 16. References

Basnet, P. (2025). *Wakil-G Chat.* Retrieved from wakilg.com: https://chat.wakilg.com/chat

Bijukchhe, A. (n.d.). *Nepali Law Padeko Dai.* Retrieved from yeschat.ai: https://www.yeschat.ai/gpts-2OToXaRSxt-Nepali-Law-Padeko-Dai

Bijukchhe, A. (n.d.). *Nepali Law Padheko Dai.* Retrieved from ChatGPT: https://chatgpt.com/g/g-2sICoA5HJ-nepali-law-padeko-dai

Medeiros, T., Medeiros, M., Azevedo, M., Silva, M., Silva, I., & Costa, D. G. (2023). Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manual. *Vehicles*, 1384-1399.

NepKanun: A RAG-Based Nepali Legal Assistant. (2025). *ACL Rolling Review.* Association for Computational Linguistics (ACL).

Patil, V. N., Mokashi, D., Patil, A., Kharade, P., Nilje, K., Kadam, A., . . . Jadhav, P. (2024). Integrating AI-Powered Chatbots and PDF Processing for Enhanced Document Management: A Full-Stack AI SaaS Approach. *Library Progress International*, 9982–9992.

Shah, S., Ryali, S., & Venkatesh, R. (2024, November 8). *Multi-Document Financial Question Answering using LLMs.* Retrieved from arXiv: https://arxiv.org/abs/2411.07264

Silno, M., Falco, M., Croce, D., & Rosso, P. (2025). Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access*, 18253 - 18276.