NyayaNet – Comprehensive Indian Legal AI Ecosystem

Abstract / Project Description

NyayaNet is an AI-driven platform designed to modernize the Indian legal ecosystem by integrating **legal research, AI-powered automation, case prediction, document drafting, and professional networking** into a unified system.

The platform will support:

1. **Type 1 – Judgment Prediction AI:** Predict probable outcomes of cases across all Indian legal domains with probability scores.
2. **Type 2 – Legal Retrieval AI:** Retrieve relevant statutes, precedents, and arguments from millions of Indian legal documents.
3. **Type 3 – Document Drafting AI:** Automate creation of legal documents, including petitions, affidavits, contracts, and notices.
4. **Professional Social Network:** Lawyer-centric networking inspired by Instagram, LinkedIn, and Quora; includes profiles, discussion forums, Q&A, and case debates.

Objective:
Reduce time, effort, and cost in legal research and drafting while fostering collaboration and knowledge-sharing among Indian legal professionals.

Key Features:

- Comprehensive AI Assistance across prediction, retrieval, and drafting.
- Multi-domain coverage: Civil, Criminal, Consumer, Taxation, Labor, Family, Property, Corporate, and more.
- Social networking for lawyers and students, encouraging professional interaction and case discussion.
- Optional: Public guidance for legal awareness (can be added later).

Project Scope

NyayaNet is divided into **three main AI modules and a professional social network**:

Module	Purpose
Type 1 – Judgment Prediction AI	Predict the likely outcome of a case using structured case data.

Module	Purpose
Type 2 – Legal Retrieval AI	Retrieve relevant laws, precedents, and arguments based on semantic search.
Type 3 – Document Drafting AI	Automatically generate structured legal documents using case-specific data.
Professional Social Network	Profiles, discussions, Q&A, and knowledge-sharing for legal professionals.

Excluded in Initial Phase: *Public Explanation AI (Type 4) to simplify laws for citizens.*

Why Type 3 AI Requires More Dataset Work

- Unlike Type 1 and Type 2, **Type 3 AI depends on structured document templates** (petitions, contracts, affidavits).
- Existing corpora (INDLEX / Legal-BERT) **mostly contain statutes or judgment excerpts**, not fully structured templates.
- **Therefore:** Extensive **manual preprocessing and template extraction** is needed before fine-tuning the model.
- This includes identifying headings, clauses, placeholders, and mapping common structures across multiple domains.

Technical Feasibility

- **Datasets & Models**

AI Type	Dataset / Model	Pros	Challenges / Cons
Type 1 – Prediction	NyayaAnumana (702,945 preprocessed cases)	Multi-domain, structured for prediction, includes SC + HC + Tribunals	Requires GPU, some additional labeling may be needed
Type 2 – Retrieval	InLegalBERT Corpus (5.4M docs, 27 GB)	Broad coverage, pre-trained embeddings, semantic search ready	Heavy storage, preprocessing needed

AI Type	Dataset / Model	Pros	Challenges / Cons
Type 3 – Drafting	INDLEX / Legal-BERT	Indian legal language-aware, includes statutes	Needs **manual template creation and preprocessing**, not all document types present

---

- **AI Techniques & Approach**
  1. **Type 1 – Judgment Prediction**
     - Preprocessing: Extract case facts, metadata, and outcomes.
     - Fine-tune LegalBERT / LLaMA models.
     - Output: Multi-class prediction (Allowed / Dismissed / Partially Allowed / Convicted / Acquitted).
  2. **Type 2 – Legal Retrieval**
     - Preprocessing: Chunk legal documents, clean text, attach metadata.
     - Create embeddings (LegalBERT / InLegalBERT).
     - Store in vector database (FAISS recommended for offline MVP).
     - Semantic search for relevant precedents.
  3. **Type 3 – Document Drafting**
     - Manual template extraction from corpora.
     - Fine-tune LLMs to generate structured documents.
     - Requires placeholders for case-specific details.
     - Human verification recommended for legal accuracy.

AI Modules – Techniques, Workflow, and Challenges

- ***Type 1 AI – Judgment Prediction***

Objective: Predict probable outcome of a case based on its description.

Dataset: NyayaAnumana (702,945 preprocessed cases)

- Covers SC, High Courts, Tribunals, District Courts.
- Includes structured outcome labels and metadata (court, year, judges, parties).

Technique / Approach:

1. **Preprocessing**
    - Extract case facts, issues, parties, and domain (Civil / Criminal / Tax / Labor / Family / Consumer).
    - Clean text: remove headers, procedural statements, OCR artifacts.
    - Chunk very long judgments (512–1024 tokens).
    - Label alignment: Map each chunk to outcome label.
2. **Model Architecture**
    - Base: LegalBERT, LLaMA, or InLegalBERT embeddings.
    - Fine-tune on input = facts + issues → output = judgment label.
    - Multi-class classification: Allowed, Dismissed, Partially Allowed, Convicted, Acquitted.
3. **Evaluation Metrics**
    - Accuracy, Precision, Recall, F1-score.
    - Domain-specific confusion matrices.

Pros:

- Predicts likely outcomes, supports strategy planning.
- Multi-domain coverage.
- Structured dataset reduces preprocessing effort.

Challenges / Cons:

- Imbalanced labels (some case types dominate).
- Nuanced judgments may be hard to predict.
- Requires GPU/TPU for fine-tuning.

- ***Type 2 AI – Legal Retrieval***

Objective: Retrieve relevant laws, precedents, and legal arguments for a given case description.

Dataset: InLegalBERT Corpus (5.4M docs, 27GB)

Technique / Approach:

1. **Preprocessing**
    - Clean text: remove headers, OCR errors, duplicates.
    - Chunk large documents into logical segments (per article, paragraph, clause).
    - Attach metadata: case name, court, year, IPC/sections cited.
2. **Embedding Generation**
    - Use LegalBERT / InLegalBERT embeddings.
    - Store embeddings in vector DB (FAISS for offline MVP).
3. **Query Processing**
    - Convert user query into embedding.
    - Perform semantic search to fetch top-K relevant chunks.
    - Optional: RAG (Retrieval-Augmented Generation) for contextual explanations.

Pros:

- Covers all legal domains.
- Faster access to relevant precedents.
- Scalable for millions of documents.

Challenges / Cons:

- Large storage (raw + embeddings ~60–70 GB).
- Embedding generation is compute-intensive.
- Mostly English; may miss regional language judgments.

- ***Type 3 AI – Document Drafting***

Objective: Automate drafting of legal documents like petitions, affidavits, contracts, and notices.

Dataset: INDLEX / Legal-BERT corpora

Key Note: This module requires more preprocessing than others because templates are not directly present. Manual extraction of document structures is essential.

Technique / Approach:

1. **Template Extraction**

- Identify common document types across domains: petitions (family/consumer), contracts, affidavits/notices.
- Extract headings, subheadings, clauses, and placeholders (e.g., <Petitioner>, <Respondent>, <Facts>).

2. **Fine-Tuning**
    - Train LLMs (LegalBERT, LLaMA) to fill templates with case-specific details.
    - Input: case type, parties, key facts, relevant laws.
    - Output: formatted legal document.
3. **Validation**
    - Ensure generated drafts follow proper legal formatting.
    - Optional human review before submission.

Pros:

- Reduces repetitive drafting work.
- Ensures consistent terminology and formatting.
- Adaptable across domains with curated templates.

Challenges / Cons:

- Extensive dataset preprocessing required.
- Templates may not cover all document types.
- Risk of hallucination; human review is mandatory.

- ***Professional Social Network Integration***

Objective: Enable networking, discussions, Q&A, and knowledge sharing among legal professionals.

Features:

- Profiles with experience, specialization, and publications.
- Posts, discussions, and case debates (Quora + LinkedIn + Instagram style).

Pros:

- Encourages collaboration.
- Builds community engagement.

Challenges / Cons:

- Requires moderation to prevent misinformation.
- User privacy and data protection critical.

Data Preprocessing

Type 1 – Judgment Prediction AI (NyayaAnumana)

1. Extract key components: case metadata (court, year, judge, parties), case facts, outcome labels.
2. Clean text: remove headers, procedural statements, OCR artifacts.
3. Chunk long judgments (512–1024 tokens).
4. Align labels to each chunk (multi-class: Allowed, Dismissed, Partially Allowed, Convicted, Acquitted).
5. Tokenize using LegalBERT / GPT tokenizer.

Type 2 – Legal Retrieval AI (InLegalBERT)

1. Clean text: remove irrelevant metadata, OCR errors, duplicates.
2. Chunk judgments/statutes into logical sections.
3. Generate embeddings using LegalBERT / InLegalBERT.
4. Attach metadata: case name, court, year, cited sections/IPC articles.
5. Optional RAG layer for contextual explanation and summarization.

Type 3 – Document Drafting AI (INDLEX / Legal-BERT)

1. Extract templates manually for common document types.
2. Structure text into template + placeholders (e.g., <Petitioner>, <Respondent>, <Facts>, <Relevant_Section>).
3. Fine-tune LLM to fill templates using case-specific facts.
4. Validate drafts for correct formatting and legal terminology.
5. Requires **more preprocessing work** than Type 1 & 2 due to template extraction.

Storage and Hardware Recommendations

AI Module	Storage	GPU	Notes
Type 1	10–15 GB processed	16–24 GB VRAM	Fine-tuning judgment prediction models
Type 2	30–35 GB raw + 50–100 GB embeddings	Optional GPU	FAISS vector DB for semantic search
Type 3	5–10 GB	12–16 GB VRAM	Template-based fine-tuning for drafting

Assumptions for MVP:

- Storage: **OneDrive for Students (1TB)** — sufficient for datasets, embeddings, and templates.
- GPU: Provided by college.
- RAM: 64–128 GB for preprocessing and embedding generation.
- CPU: 16–32 cores for batch processing and vectorization.

Challenges and Risk Mitigation

Challenge	Risk	Mitigation
Data Privacy	Legal documents may contain sensitive info	Anonymize parties, remove PII
Bias in AI Prediction	Historical judgments reflect bias	Fine-tune carefully, use multiple sources
Dataset Size & Compute	Large corpora → slow processing	Use batch processing, cloud GPU for fine-tuning
User Trust	Lawyers may distrust AI	Show confidence scores, human verification
Law Updates	Frequent amendments	Automate scraping, schedule periodic updates
Drafting Accuracy	AI hallucination	Curated templates, mandatory human review

---

Pros and Cons of Entire Platform

Pros:

- Integrates prediction, retrieval, drafting, and networking.
- Covers all Indian legal domains.
- Reduces research and drafting time.
- Encourages knowledge-sharing among lawyers.
- Scalable for students, NGOs, and researchers.

Cons:

- High computational resources required for fine-tuning.
- Extensive dataset cleaning and preprocessing, especially for Type 3 AI.
- AI prediction (Type 1) may reflect historical biases.
- Drafting AI depends on curated templates; risk of hallucination.
- Public Explanation AI (Type 4) not included in initial phase.

Cost Estimate (Minimal Setup)

Component	Minimal Setup Approach	Estimated Cost
Datasets	NyayaAnumana, InLegalBERT, INDLEX / Legal-BERT	₹0
GPU Compute	College-provided GPU	₹0
Vector DB	FAISS local, open-source	₹0
Storage	OneDrive for Students (1TB)	₹0
Miscellaneous Tools	Optional scripts, preprocessing tools	₹500–1,000 one-time

Tip: Start MVP with Type 2 + Social Network locally, then integrate Type 1 & 3 gradually.

Conclusion & Expected Outcomes

NyayaNet aims to create a **holistic AI-powered legal ecosystem** integrating:

1. Judgment Prediction AI (Type 1)
2. Legal Retrieval AI (Type 2)
3. Document Drafting AI (Type 3)
4. Lawyer-focused Social Network

For Lawyers:

- Faster access to precedents and laws.
- AI-assisted draft documents.
- Probabilistic case outcome prediction to aid strategy.

For Law Students & Researchers:

- Centralized learning resources.
- Semantic search across judgments and laws.
- Exposure to real-world legal drafting and cases.

For Legal Community:

- Networking platform for discussion and collaboration.
- Community-driven Q&A forums.
- Increased cross-domain collaboration between lawyers.

MVP / Initial Version:

- Start with **Type 2 AI + Social Network**.
- Gradually integrate **Type 1 & Type 3 AI**.

Long-Term Vision:

- Full-fledged AI ecosystem for all Indian legal domains.
- Optional Type 4 AI for public-friendly legal explanations in simple language.