LEGAL SYSTEM: CASE BRIDGE

Unified Legal Retrieval and Generation

T AKSHAYA G474

SCOPE OF THE PROJECT

- **Dataset Utilization:** Use of the CLERC dataset for legal case retrieval (IR) and retrieval-augmented generation (RAG).
- **Model Development:** Fine-tuning domain-specific retrieval models (e.g., LegalBERT DPR) for improved citation recall and precision
- **Legal Analysis Generation:** Generating accurate legal analyses with minimal hallucination, using retrieved relevant cases.
- **Evaluation:** Measuring performance using Recall, ROUGE, Citation Precision/Recall, and False Positive metrics.

Basic Understanding of the Base Paper

- CLERC is a large-scale dataset designed for training/evaluating AI models on:
  - Information Retrieval (IR): Find relevant legal cases.
  - Retrieval-Augmented Generation (RAG): Generate analytical legal text with accurate citations.
- Built on U.S. federal case law (CAP corpus, 1.84M documents).
- Enables models to co-write legal analyses and assist lawyers in referencing precise cases.

Dataset Structure:

- **CLERC/doc** – Full case documents.
- **CLERC/passage** – Case documents split into 350-word passages.
- **CLERC/generation** – Passages for testing legal analysis generation.

Business Problem Addressed

- Lawyers spend significant time/effort searching for relevant precedents.
- Existing solutions (Westlaw, LexisNexis):
  - Expensive, closed-source, manual annotation.
  - Lack of automation and scalability.
- Business Problem: Inefficient, time-consuming, costly legal research and analysis.
- General AI tools may generate inaccurate or unsupported claims without proper grounding in actual legal texts.

Key Insights from the CLERC Research Paper

Data Introduction

CLERC is a large dataset of U.S. federal court cases, meticulously curated for AI training.

Dual Task Enablement

It supports both relevant case citation retrieval and legal analysis generation from those citations.

Challenging Benchmark

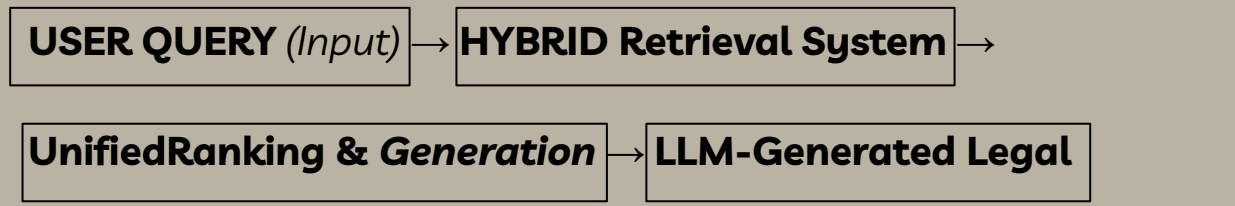
Queries are designed to mimic lawyer practices, testing **long-context reasoning, citation precision**, and hallucination reduction

Driving Domain-Specific AI

Serves as a standardized benchmark to advance accurate, trustworthy, and domain-tuned legal AI systems.

Proposed Solution: Hybrid Retrieval + AI Legal Analysis

- We combine a hybrid legal case retrieval pipeline with advanced AI-powered analysis to enhance legal research speed, accuracy, and reliability.



Hybrid Retrieval

- Uses multiple retrieval models (keyword-based semantic) for legal coverage

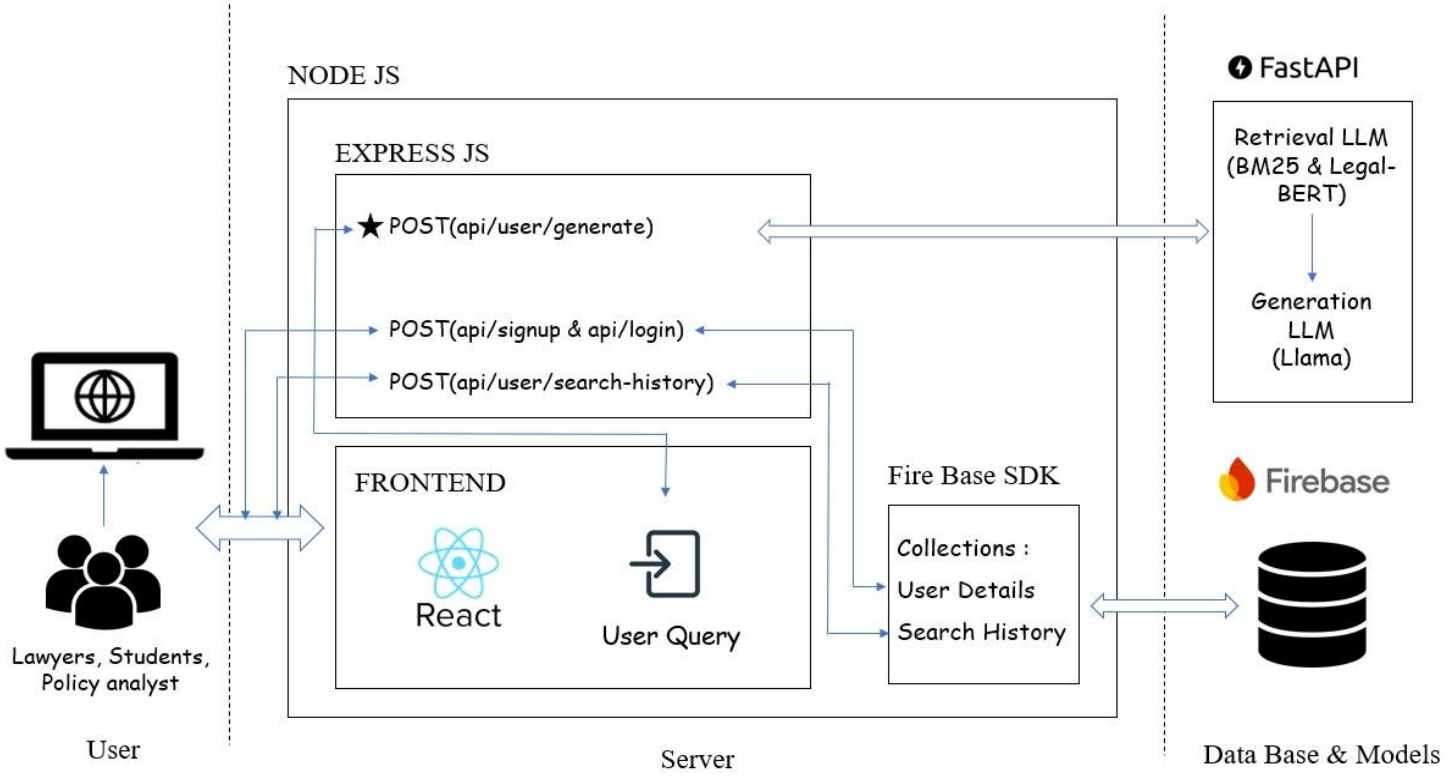
Unified Ranking & Generation

Filters, ranks, and selects the most relevant legal references.

LLM-Generated Legal Brief

Produces a structured, comprehensive citation-backed legal brief for easy interpretation.

Basic Architecture Diagram



RESEARCH PAPER IDEA –

Indian Legal AI Dataset (Inspired by CLERC)

The Idea

Create a **CLERC-style benchmark** for the Indian legal system, built from Supreme Court & High Court case laws.

Why It Matters

Indian legal professionals need **fast, accurate case retrieval** and **AI-assisted legal analysis**, just like U.S. lawyers benefit from CLERC.

How It Could Work

Collect preprocess and digitize. Adapt CLERC's dual-task design, and follow Indian citation standards.

The Vision

A **multilingual, India-focused legal AI dataset** that becomes the gold standard for legal AI research in India

THANK YOU!