Pavithra Manikandan: 38819531
Navya Saxena: 52078368
Aniket Ghorpade: 52552619
John Harshith Kavuturu: 67055516

Context-Aware Legal Information Retrieval

Motivation:

Judges, attorneys, and researchers spend a lot of time hunting through huge databases to find the right cases and opinions. Most search tools rely heavily on exact keywords or citation matching, which often fails to capture the many ways the same legal idea can be phrased.

For example:
Query: “Cases recognizing that online agreements can be enforceable.”
Relevant passage: “Click-wrap contracts were upheld as binding by the court.”

Both say the same thing, but the wording is different. A keyword search might miss that “online agreements” ≈ “click-wrap contracts.” Legal writing is dense, context-heavy, and full of references to prior rulings, so understanding meaning and context, not just words, is essential. Newer semantic models (e.g., Sentence-BERT, DPR, cross-encoders) can match on meaning, but they’ve mostly been tested on general datasets (like MS MARCO or Wikipedia), not in law, where reasoning and interpretation are central. Our goal is to test whether context-aware retrieval can make it easier to surface the right legal passages by incorporating semantic and contextual understanding into the retrieval process.

Project Goal:

The primary objective of our project is to design and evaluate a context-aware legal retrieval system that identifies relevant case-law passages based on their conceptual and contextual relationship to a query, rather than on simple lexical overlap. We’ll answer two questions:

1. How well do semantic retrieval models capture legal relevance compared with traditional keyword methods?
2. Does adding extra context, via reranking or query expansion, improve results on legal texts?

Approach and Data:

We plan to utilize the LePaRD (Legal Passage Retrieval Dataset) as the main dataset, supplemented by LegalBench-RAG for cross-domain validation. Both datasets contain natural-language legal queries paired with annotated relevant passages from U.S. court opinions and statutes.

Implementation plan

1. Baseline (lexical): BM25 keyword retriever.
2. Strong baseline (semantic): Sentence-BERT dense retriever to capture meaning.
3. Extensions:
   a. Cross-encoder reranker: jointly encodes query + passage to model interactions and legal reasoning.
   b. Query expansion: use google/flan-t5-base to rewrite/paraphrase queries (e.g., “precedent” → “case law”) to broaden recall.

Performance will be measured using standard information-retrieval metrics such as Recall@10, Mean Reciprocal Rank (MRR), and nDCG, which evaluate both the ranking quality and coverage of retrieved results.

Expected Outcomes:

- A head-to-head comparison of keyword, semantic, and context-aware methods for legal search.
- Evidence of how contextual modeling changes retrieval accuracy and how neural models handle complex legal language.
- A reproducible codebase, metric dashboards, and visual examples of retrieved results.

Deliverables will include a reproducible codebase, quantitative performance reports, and visual demonstrations of retrieval outcomes. Ultimately, we aim to show how modern language models can make legal research tools smarter and more “meaning-first,” helping practitioners and researchers find the right case law faster.