Group ID: 18

LegalMind: *Structured Retrieval and Reranking for Transparent Legal Question-Answering*

Abstract

Legal information retrieval presents unique challenges that make it an ideal domain for investigating advanced retrieval techniques. Legal professionals spend a significant portion of their workday - estimates suggest 30-40% spent searching through vast document collections to find relevant case law, statutes, and regulations. This process is time-intensive, expensive, and often relies on keyword searches that may miss conceptually relevant information.

The complexity of legal language creates particular difficulties for retrieval systems. Legal documents contain specialized terminology, complex sentence structures, numerous cross-references, and concepts that require domain knowledge to properly interpret. A simple term like "consideration" carries specific meaning in contract law that differs from everyday usage. Additionally, legal documents tend to be lengthy, with relevant information often distributed across multiple sections or documents.

These challenges present an opportunity to evaluate whether modern retrieval techniques can meaningfully improve information access in this specialized domain compared to traditional approaches. Effective legal retrieval systems could potentially reduce research time, lower costs, and ultimately improve access to justice.

Project Approach

This project aims to systematically compare different retrieval strategies for legal question answering. Rather than simply implementing a single approach, we will evaluate multiple retrieval architectures against the same legal dataset, allowing for direct performance comparisons. This comparative analysis will provide insights into which techniques are most effective for the specific challenges presented by legal text.

Experimental Design

We have designed four experimental configurations to test progressively more sophisticated retrieval approaches:

Experiment 1: Lexical Retrieval Baseline

The foundation is a traditional retrieval system using BM25, a proven probabilistic ranking function widely used in search engines. This approach scores documents based on term frequency, inverse document frequency, and document length normalization. It essentially measures overlap between query terms and document terms, with adjustments for document length and term importance. This provides a strong, industry-standard baseline against which to measure improvements.

Experiment 2: Dense Retrieval with Legal Language Model

The second configuration employs dense retrieval using a transformer-based language model pre-trained or fine-tuned on legal corpora. This approach encodes both queries and documents as dense vector representations, capturing semantic relationships beyond exact keyword matching. Documents are retrieved based on vector similarity (cosine similarity) between the query embedding and document embeddings. This method potentially captures conceptual relationships that lexical retrieval might miss.

Experiment 3: BM25 Retrieval with Cross-Encoder Reranking

A two-stage retrieval pipeline combining the efficiency of lexical search with neural reranking precision. BM25 generates an initial candidate set (top 50 documents), which are then passed to a cross-encoder model that processes each query-document pair jointly, capturing complex interactions between terms. This approach leverages BM25's efficiency for broad candidate generation while applying the resource-intensive neural reranker only to promising documents.

Experiment 4: Dense Retrieval with Cross-Encoder Reranking

This configuration applies the same cross-encoder reranking to documents retrieved by the dense retriever. This design examines whether semantic understanding in dense retrieval can be further enhanced by query-document interaction modeling, and allows comparison with Experiment 3 to determine how the first-stage retriever impacts reranking effectiveness.

Literature Review

This project builds upon key advancements in retrieval-based Question Answering (QA), with a focus on the legal domain.

Retrieval Approaches: Traditional sparse retrieval techniques like BM25 have been foundational in information retrieval. Khazaeli et al. (2021) developed a legal QA system combining sparse vector search with BERT-based reranking. To address lexical matching limitations, dense retrieval models encode semantic meaning in vector representations. Louis et al. (2023) proposed a graph-augmented dense statute retriever (G-DSR) incorporating legislation structure via graph neural networks.

Reranking Techniques: Reranking has emerged as a crucial component in modern retrieval pipelines. Askari et al. (2022) introduced a three-stage framework comprising pre-training, fine-tuning, and reranking to enhance retrieval using contextual similarity. Althammer et al. (2021) explored summarization-based reranking for case law retrieval, combining lexical and dense methods with BERT-based rerankers fine-tuned on case summaries.

Challenges in Legal Information Retrieval: Legal texts present unique challenges including specialized terminology, complex document structures with long-range dependencies, and evolving precedents - factors necessitating specialized approaches beyond general-domain techniques.

Group ID: 18

Implementation Progress

We are currently in the initial phases of the project, with our primary focus on gathering and preparing the COLIEE + CaseHOLD datasets. At this stage, we have:

1. Established the project requirements and experimental design
2. Identified the necessary libraries and frameworks for implementation
3. Begun the data collection process for the COLIEE corpus and CaseHOLD corpus
4. Created a preliminary data processing pipeline design

No actual implementation of the retrievers or re-rankers has commenced yet, as our current priority is ensuring proper dataset acquisition and preparation. Once we have secured the complete dataset, we will proceed with implementing the BM25 baseline system using Pyserini, followed by the dense retriever and reranking components.

Our implementation timeline accounts for this staged approach, with dataset preparation serving as the critical first step before any retrieval system development can begin. We anticipate starting the implementation of the BM25 baseline immediately after completing the dataset preparation phase.

Plan for Next Steps

Our project roadmap prioritizes the following key phases:

1. **Dataset Acquisition and Preparation (2 weeks)**
   - Gather COLIEE & CaseHOLD datasets, perform cleaning, and create validation splits
2. **Retriever Implementation (3 weeks)**
   - Implement BM25 baseline with Pyserini
   - Develop dense retriever with legal domain language model
   - Configure indexing and query processing for both approaches
3. **Reranking and Hybrid Pipeline Development (2 weeks)**
   - Implement cross-encoder reranking module
   - Integrate with both retrieval systems (BM25 + reranker, Dense + reranker)
4. **Evaluation and Analysis (2 weeks)**
   - Conduct comparative performance analysis using IR metrics
   - Perform error analysis and qualitative assessment
   - Identify strengths/weaknesses of each approach for legal questions
5. **Final Documentation (1 week)**
   - Compile findings, implementation details, and visualizations.