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

Context-Aware Legal Information Retrieval

Dataset:

Training Dataset: LePaRD https://huggingface.co/datasets/rmahari/LePaRD

```
from datasets import load_dataset
ds = load_dataset("rmahari/LePaRD")

DatasetDict({
    train: Dataset({
                features: ['dest_id', 'source_id', 'dest_date', 'dest_court', 'dest_name', 'dest_cite', 'source_date', 'source_court', 'source_name', 'source_cite', 'passage_id', 'quote', 'destination_context'],
        num_rows: 22744882
    })
})
```

Testing benchmark: LegalBench-RAG
https://www.dropbox.com/scl/fo/r7xfa5i3hdsbxex1w6amw/AID389Olvtm-ZLTKAPrw6k4?rlkey=5n8zrbk4c08lbit3iiexofmwg&e=1&st=0hu354cq&dl=0

For this project, we selected the LePaRD (Legal Passage Retrieval Dataset) as our primary training data. LePaRD contains over 22 million examples drawn from U.S. federal court decisions, where a legal opinion cites another case and quotes a specific passage. Each example connects a query-like quote to its destination context, the surrounding text in the cited case, making it ideal for training a supervised retrieval model. This structure closely matches our task: given a legal question or argument, retrieve the relevant passage from case law that supports or addresses it.

For evaluation, we use the LegalBench-RAG benchmark, a curated dataset designed to test retrieval systems on challenging, natural-language legal queries. Unlike LePaRD, which is large and citation-driven, LegalBench-RAG focuses on short-answer retrieval, requiring the system to match a plain-text query with its relevant case passage. This setup enables us to evaluate how well our trained system generalizes to real-world legal search tasks that go beyond citation matching.

We chose these datasets because they reflect both the scale and specificity of real legal research. LePaRD allows us to train dense retrievers and rerankers using actual judicial citation behavior, while LegalBench-RAG gives us a domain-relevant testbed for measuring effectiveness under realistic usage scenarios.

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

Literature Review:

The ASKE paper [1] proposes a context-aware, unsupervised framework for the extraction of legal concepts and classification of legal documents by leveraging large language model embeddings and zero-shot learning. Unlike traditional, rule-based, or supervised approaches, ASKE builds a conceptual graph of legal topics through iterative extraction, enrichment, and refinement of terms from large legal corpora. It represents each legal concept as a semantic embedding cluster and evaluates its quality against cosine similarity to EuroVoc label embeddings. Its pipeline is organized in three phases: an initial keyword seeding phase, semantic expansion by embeddings, and iterative graph-based refinement. Most important is that ASKE works without labeled training data and thus is perfectly fit for resource-scarce legal domains where supervised annotations are poor or inconsistent.

ASKE was evaluated on Eur-Lex using pseudo-precision and pseudo-recall computed via FastText-based semantic similarity. For concept extraction, it performed better compared with the baseline models like ZSTM and BERTopic, achieving a pseudo-precision of 0.831 and pseudo-recall of 0.811 at generation 21, while baselines achieved less than 0.83 and 0.79, respectively. It was able to achieve an F1-equivalent of 0.625, outperforming all other models, including ZSTM@2000 and BERTopic, during document classification at generation 15. These results were statistically significant. These results show that ASKE's context aware generation based approach successfully captures abstract legal concepts in context and outperforms the state-of-the-art baselines, without the need for pre-defined numbers of topics. While ASKE focuses on the task of legal information extraction and conceptual classification rather than document retrieval, its use of context-aware embeddings has direct relevance for our project. It shows that embedding-based representations can effectively capture the semantic meaning of legal phrases and categories. Nevertheless, this system is optimized for concept discovery and classification, not full passage-level retrieval based on complex queries. Our work builds on a similar motivation of capturing legal meaning beyond keywords, but instead targets passage level retrieval using fine-tuned semantic retrievers and cross-encoders. This places our system closer to the practical task of helping legal researchers find relevant case law.

The SAILER paper [2] addresses a critical challenge in legal case retrieval: while pre-trained language models have achieved success in general ad-hoc retrieval tasks, they struggle to capture the unique characteristics of legal documents. Legal cases present two distinct challenges that existing models fail to handle effectively. First, legal documents are lengthy texts (often thousands of words) with intrinsic logical structures divided into five standard sections: Procedure, Fact, Reasoning, Decision, and Tail. Most existing language models either have limited capacity for long documents or ignore these structural dependencies entirely. Second, legal relevance is highly sensitive to key legal elements, even subtle differences in specific legal circumstances can lead to completely different judgments. For example, two paragraphs that are nearly identical in wording may be legally irrelevant if one describes “stealing” while the other describes “purchasing” property. To address these issues, SAILER proposes a structure-aware pre-training framework that explicitly models the logical relationships between

different sections of legal documents. The model uses an asymmetric encoder-decoder architecture where a deep encoder processes the Fact section into dense vectors, and two shallow decoders reconstruct aggressively-masked text from the Reasoning and Decision sections using the Fact representation. This design simulates how legal experts write judgments: they analyze facts, identify relevant legal elements in their reasoning, and render decisions based on those elements.

SAILER was evaluated on four legal case retrieval benchmarks spanning both Chinese (LeCaRD, CAIL2022-LCR) and English (COLIEE2020, COLIEE2021) legal systems in both zero-shot and fine-tuning settings. In zero-shot evaluation on LeCaRD, SAILER achieved an NDCG@30 of 0.8485, significantly outperforming retrieval-oriented baselines like RetroMAE (0.7089) and even surpassing traditional BM25 (0.8172), making it the only neural model to beat lexical matching without supervised data. On CAIL2022-LCR, it achieved 0.8660 NDCG@30 compared to 0.8545 for BM25. In fine-tuning experiments, SAILER achieved the best performance across all metrics on both COLIEE datasets, reaching 0.6164 F1-score on COLIEE2020 and 0.5298 Recall@100 on COLIEE2021, demonstrating its ability to distinguish relevant cases even in large candidate pools. Ablation studies confirmed that both the Reasoning Decoder and Decision Decoder contribute significantly to performance, with removing both causing NDCG@10 to drop from 0.7979 to 0.6479. Visualization analysis showed SAILER's vectors could accurately discriminate between six easily-confused criminal charges without any supervised data, while baseline models like BERT mixed different charges together. This research directly validates the importance of context-aware modeling for legal retrieval, which motivates our project's approach. While SAILER focuses on exploiting document structure through encoder-decoder pre-training, our work leverages a complementary signal: the citation structure inherent in LePaRD, where cases cite and quote other cases. Both approaches recognize that legal documents contain rich structural signals beyond raw text. However, SAILER's limitation is that it requires parsing documents into their component sections, which may not always be reliable across different jurisdictions or document formats. Our citation-based approach on LePaRD may offer a more generalizable training signal since legal citations are universally formatted and explicitly marked in case law. Furthermore, our project extends beyond retrieval by incorporating cross-encoder reranking and query expansion to bridge the semantic gap between how users phrase queries and how relevant legal concepts appear in case law by directly addressing the “online agreements” versus “click-wrap contracts” challenge that motivates our work.

The paper [3] presents LegalQA, a benchmark of 9,846 consumer questions and 33,670 lawyer-written answers, and introduces CEFS, a cross-encoder reranker that preserves a question’s structure by marking the subject [S], description [D], and tags [T] in its input. Built as a two-stage pipeline, BM25 for candidate gathering, then CEFS for reranking, it yields strong gains over competitive cross-encoder baselines, raising MAP@1k to 0.270 (vs. 0.236 LegalQA-tuned and 0.109 MS MARCO-tuned). Recall also improves (R@10 = 0.428, R@1 = 0.209), and ablations show the description contributes most, then the subject, with tags adding a smaller boost. The results highlight a real lexical gap between lay phrasing and attorney language, which structure-aware reranking helps bridge.

However, CEFS still depends on a lexical first stage and on question-side structure (especially tags) that may be missing or inconsistent, and it is evaluated on a narrow community bankruptcy domain rather than passage-level case law. It also adds latency and cannot recover items BM25 never retrieved. Our project addresses these limits by using semantic first-stage retrieval (e.g., Sentence-BERT/DPR) with query expansion to lift recall, then applying a cross-encoder reranker to passage-level case law (LePaRD, LegalBench-RAG), aiming to combine structure-aware reasoning with broader coverage of the sources practitioners actually search.

Citations:

[1] Castano, S., Ferrara, A., Furiosi, E., Montanelli, S., Picascia, S., Riva, D., & Stefanetti, C. (2024). Enforcing legal information extraction through context-aware techniques: The ASKE approach. Computer Law & Security Review, 52, 105903. https://www.sciencedirect.com/science/article/pii/S0267364923001139

[2] Li, H., Ai, Q., Chen, J., Dong, Q., Wu, Y., Liu, Y., Chen, C., & Tian, Q. (2023). SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. https://arxiv.org/abs/2304.11370

[3] Askari, Arian, et al. "Answer retrieval in legal community question answering." European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2024. https://arxiv.org/abs/2401.04852