CIS5300 Milestone 3: Hybrid Lexical + Semantic Retrieval
Project: Context-Aware Legal Information Retrieval

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

November 25, 2025

1 Introduction

This report describes our Milestone 3 extension: a hybrid retrieval system that combines BM25 lexical retrieval with Sentence-BERT semantic retrieval. Building on Milestone 2's baselines (TF-IDF, BM25, and Sentence-BERT), we demonstrate that fusing complementary retrieval signals significantly improves performance on the legal passage retrieval task.

2 Extension: Hybrid BM25 + Sentence-BERT Retrieval

2.1 Motivation

Milestone 2 established that BM25 excels at matching exact legal terminology but fails to capture semantic relationships and paraphrases, while Sentence-BERT captures semantic similarity but underperforms on rare legal terms requiring exact lexical matching. To bridge this gap, we implemented a *hybrid retriever* that combines both approaches through weighted score fusion, leveraging the complementary strengths of lexical and semantic matching.

2.2 Method

Our hybrid system indexes each 500-word document chunk using both BM25 (with standard parameters $k_1 = 1.5$, $b = 0.75$) and Sentence-BERT embeddings (using `all-mpnet-base-v2`). At query time, we compute normalized scores from both retrievers and combine them using a weighted sum:

$$S_{\text{hybrid}} = \alpha \cdot S_{\text{BM25}} + (1 - \alpha) \cdot S_{\text{SBERT}} \tag{1}$$

where S_{BM25} and S_{SBERT} are min-max normalized to [0,1] before fusion, and $\alpha = 0.55$ (tuned to maximize Recall@10). We consider the top-100 candidates from each retriever (fusion depth) before combining scores. The entire pipeline remains training-free and efficient, requiring only pretrained Sentence-BERT embeddings and standard BM25 indexing.

2.3 Implementation

- **Fusion method:** Weighted combination of normalized scores
- **Weights:** BM25 = 0.55, Sentence-BERT = 0.45 (tuned on validation set)

- **Fusion depth:** Top-100 candidates per retriever
- **Model:** `sentence-transformers/all-mpnet-base-v2`
- **Batch size:** 32 passages per encoding batch

3 Empirical Evaluation

3.1 Dataset

We evaluate on the full LegalBench-RAG ContractNLI benchmark:

- Test set: 977 queries from ContractNLI
- Corpus: Legal contract documents (563 passages after chunking)
- Evaluation metrics: Exact Match, Span F1, Recall@10, nDCG@10

3.2 Results

Table 1 shows that the hybrid approach substantially outperforms all individual baselines. The hybrid retriever achieves Recall@10 = 0.5511, representing a +3.7 point improvement over BM25 (0.5137) and a +12.0 point improvement over Sentence-BERT alone (0.4317). Similarly, nDCG@10 = 0.4808 improves by +3.6 points over BM25 and +15.8 points over Sentence-BERT. The Span F1 score also improves to 0.2357, indicating better passage-level relevance.

Table 1: Retrieval performance on ContractNLI (977 queries, $k = 10$)

Model	**Exact Match**	**Span F1**	**Recall@10**	**nDCG@10**
TF-IDF (simple baseline)	0.0000	0.2018	0.3090	0.2204
BM25 (strong baseline)	0.0000	0.2315	0.5137	0.4445
Sentence-BERT (dense)	0.0000	0.2147	0.4317	0.3232
Hybrid (BM25 + SBERT)	**0.0000**	**0.2357**	**0.5511**	**0.4808**

3.3 Observations

- The hybrid approach demonstrates that combining complementary retrieval signals (lexical and semantic) produces more reliable results than relying solely on either approach. BM25's strength in exact term matching complements Sentence-BERT's ability to capture semantic relationships and paraphrases.
- The hybrid method achieves a 7.3% relative improvement in Recall@10 over BM25 (0.5511 vs 0.5137) and a 27.8% improvement over Sentence-BERT (0.5511 vs 0.4317), confirming that both signals contribute meaningfully to retrieval quality.
- All methods achieve 0.0000 Exact Match because ContractNLI gold spans rarely align exactly with 500-word chunks. Future work could incorporate passage trimming or cross-encoder reranking to address this limitation.
- The improvement in nDCG@10 (0.4808 vs 0.4445 for BM25) indicates that the hybrid approach not only retrieves more relevant passages but also ranks them better, with relevant items appearing earlier in the ranked list.

4 Conclusion

We have successfully implemented a hybrid retrieval system that combines BM25 lexical retrieval with Sentence-BERT semantic retrieval through weighted score fusion. On the full ContractNLI benchmark (977 queries), the hybrid approach achieves Recall@10 = 0.5511 and nDCG@10 = 0.4808, outperforming both individual baselines. This demonstrates that ensemble methods combining complementary retrieval signals yield more robust performance than relying solely on lexical or semantic matching. The hybrid method leverages BM25's strength in exact term matching while preserving Sentence-BERT's ability to capture semantic relationships, resulting in improved recall and ranking quality. This work reinforces key course concepts: ensemble methods, dense/sparse hybrid retrieval, and the importance of evaluation metrics in information retrieval systems.

Code and Repository

All code for this extension, including the hybrid retriever implementation (`hybrid-baseline.py`), documentation (`hybrid-baseline.md`), as well as the Google colab notebook (`Milestone3.ipynb`), is available in our project repository:

https://github.com/JohnHarshith/CIS5300-F2025-Project

You can check all of the work done for Milestone 3, including the hybrid baseline implementation, evaluation scripts, and related notebooks, in the repository above.

Appendix

Several promising directions remain to further improve legal passage retrieval performance. First, addressing the zero Exact Match scores requires better alignment between retrieved chunks and gold answer spans. We plan to implement a two-stage pipeline: (1) use the hybrid retriever to fetch top-k candidates (e.g., $k = 100$), then (2) apply a cross-encoder re-ranker (e.g., a fine-tuned BERT model) that scores query-passage pairs to push exact gold spans toward the top. Additionally, we will experiment with dynamic passage trimming that extracts sub-spans from retrieved chunks using span prediction models, similar to approaches in question answering systems. Second, domain-specific adaptation could significantly boost semantic retrieval. We will fine-tune the Sentence-BERT encoder on legal domain data using contrastive learning on ContractNLI question-answer pairs, or leverage legal-domain pretrained models (e.g., Legal-BERT) as initialization. Third, scaling to larger corpora requires approximate nearest neighbor (ANN) indexing: we plan to integrate FAISS for efficient dense retrieval over millions of passages, enabling evaluation on the full CUAD and ContractNLI corpora without the current 10K passage limit. Finally, we will explore advanced fusion strategies beyond weighted combination, including learned fusion weights via neural networks, Reciprocal Rank Fusion (RRF) with tuned parameters, and query-dependent fusion that adapts weights based on query characteristics (e.g., technical legal terms vs. paraphrased questions). We plan to implement and evaluate these extensions in Milestone 4 as time permits to further improve retrieval performance on the legal passage retrieval task.