Bilingual Hybrid Judicial-Opinion Retrieval for Puerto Rico and Texas: Dense, Lexical, and LLM-Augmented

Andres J. Pérez Cabezas
Advisor: Lisabel Rodríguez Espinosa, J.D.
Master in Computer Science
Graduate Project EXPO, May 2026

Abstract

We present a bilingual (EN/ES) IR prototype over judicial opinions from Puerto Rico and Texas courts. Opinions from the CourtListener API are cleaned, paragraph-aware chunked, embedded with a multilingual sentence-transformer, and indexed in FAISS; a parallel BM25 index supports lexical retrieval. A weighted Reciprocal Rank Fusion (RRF) layer blends both rankings, and an optional layer adds query rewriting and cross-encoder reranking via a locally hosted Gemma 3 LLM. On a stratified 32-query held-out split, hybrid RRF reaches nDCG@10 = 0.5615; reranking raises this to 0.7680 at ~20× the latency.

Introduction

Legal research over judicial opinions is fundamental for attorneys, clerks, and academic researchers, yet existing tools for genuinely bilingual jurisdictions remain limited. Puerto Rico courts publish predominantly in Spanish; Texas courts almost exclusively in English. Commercial platforms rarely expose the retrieval mechanism, intermediate scores, or evaluation protocol, making it difficult to characterize where and why a system fails on a given query. This work addresses the transparency and bilingual-evaluation gap with a reproducible prototype that compares dense, lexical, hybrid, and LLM-augmented retrieval under a fixed protocol.

Background

Legal IR has matured along two largely separate tracks — sparse lexical scoring and dense neural embeddings — rarely combined, with bilingual evaluation scarce:

- **TREC Legal Track** set the standard evaluation paradigm for English legal retrieval; corpora and queries are monolingual.
- **BM25** (Robertson & Zaragoza, 2009) is a strong term-overlap baseline but ignores semantics and cross-language synonymy.
- **ColBERT and DPR** show dense retrievers outperform BM25 on general benchmarks, but not designed for legal or mixed-language text.
- **mBERT & multilingual MiniLM** enable cross-lingual transfer; evaluations target newswire or European pairs — not Spanish/English legal text.
- **Reciprocal Rank Fusion** (Cormack et al., 2009) combines independent rankers without learned weights.
- **LLM rerankers** (Nogueira & Cho, 2019) raise ranking quality at substantial latency cost.

Bilingual legal IR is harder: jurisdiction-specific terminology, diacritics whitespace BM25 cannot normalize, and concepts expressed differently across languages.

Problem

Construct a reproducible bilingual retrieval prototype over Puerto Rico and Texas judicial opinions that (i) integrates dense, lexical, hybrid, and optional LLM-augmented ranking, (ii) is evaluated on a stratified held-out query set with held-fixed hyperparameters, and (iii) preserves end-to-end chunk-level traceability from query to source opinion.

Methodology

Corpus and ingestion
10 court sources from the CourtListener API, capped at 2,000 opinions per court (1 req/s rate limit). Final corpus: 20,128 cleaned opinions in total from 1990–2026, ~6,000 Puerto Rico (Spanish-dominant) and ~14,000 Texas (English-dominant).

Cleaning and chunking
HTML stripping, Unicode-mojibake repair via ftfy, whitespace normalization. Opinions are split into paragraph-aware overlapping chunks (target 512 tokens, overlap 64, min 50), producing 351,564 chunks (~17 chunks/opinion).

Indexes
- **Dense:** paraphrase-multilingual-MiniLM-L12-v2 → 384-d L2-normalized vectors → FAISS IndexFlatIP (exact cosine under unit norm).
- **Lexical:** rank-bm25, whitespace tokenization, shared chunk-level metadata schema.

Hybrid scoring (weighted RRF)
Each retriever output is deduplicated at the opinion level (first-seen rank order) before fusion. For a candidate d with semantic rank rs and BM25 rank rb:

$$Score(d) = \frac{ws \cdot 1}{k + rs} + \frac{wb \cdot 1}{k + rb}$$

with ws = 0.7, wb = 0.3, k = 30 (tune-split values).

Optional LLM layer
Gemma 3 12B (FP16) via local Ollama. Query rewriter expands the query with EN/ES synonyms and legal terminology; cross-encoder reranker scores 50 query–chunk pairs on a 0–10 scale and re-sorts. Prompt outputs are SHA-256 cached; on provider failure both stages degrade to plain hybrid.

Evaluation protocol
100 manually curated queries balanced 50/50 across jurisdiction (PR/TX) and language (EN/ES). Stratified tune/held-out split (seed 42) yields 32 held-out queries 8 per stratum. Opinion-level metrics: P@10, R@10, MRR, nDCG@10 (primary). All hyperparameters frozen on the tune split; reported figures come exclusively from the held-out split.

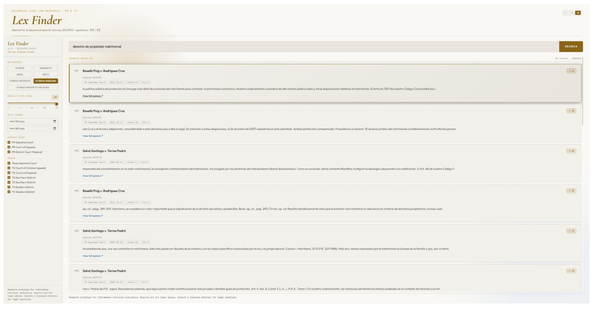

***Figure 1**. Weighted hybrid retrieval with opinion-level deduplication.*

Results and Discussion

Retriever	P@10	R@10	MRR	nDCG@10	Latency
BM25	0.381	0.241	0.584	0.418	< 1 s
Semantic (FAISS)	0.481	0.246	0.784	0.549	< 1 s
Hybrid RRF	**0.488**	**0.265**	**0.781**	**0.562**	**~2.6 s**
Hybrid + rewrite	0.444	0.247	0.703	0.518	~8 s
Hybrid + rerank	**0.594**	**0.323**	**0.901**	**0.768**	**~52 s**
Hybrid + rew. + rer.	0.547	0.297	0.852	0.706	~60 s

***Table 1**. Held-out retrieval performance (n = 32 queries).*

P@10 fraction of top-10 results that are relevant.
R@10 fraction of all relevant opinions found in the top 10.
MRR Mean Reciprocal Rank: how close to the top the first relevant result appears.
nDCG@10 ranking quality weighted by position (0–1, higher is better). Primary metric.
Latency mean wall-clock time per query on the test workstation.

Findings
- **Weighted hybrid RRF** improves over BM25 on every metric and over semantic on P@10, R@10, and nDCG@10 operational default.
- **Cross-encoder reranking** lifts nDCG@10 by +0.207 absolute over hybrid but is dominated by ~50 sequential LLM prompt calls per query.
- **Bilingual query rewriting** underperforms the hybrid baseline synonym expansion drifts BM25 away from term overlap and the dense embedding toward generic legal vocabulary.

Reproducibility
Two evaluation-affecting bugs were uncovered and pinned by real-data regression tests: missing opinion-level dedup before RRF, and cp1252 decoding of split/query files corrupting accented Spanish terms.

Hardware: AMD Ryzen 9 7950X3D, 64 GB RAM, NVIDIA RTX 4090 (24 GB), gemma3:12b FP16 via Ollama. FAISS ≈ 540 MB; BM25 ≈ 1.2 GB.

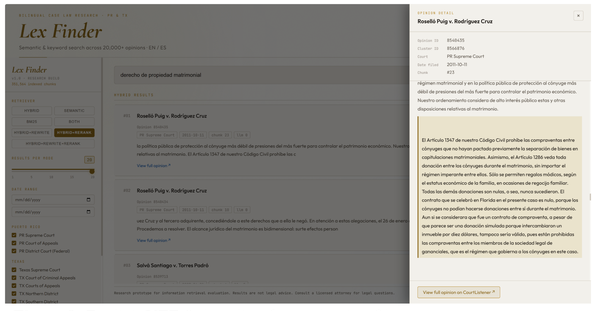

***Figure 2**. Explicit UTF-8 decoding for evaluation files.*

Conclusions

- **Hybrid search wins in practice.** Combining meaning-based and keyword search gives the best results at fast, interactive speeds (~2.6 s per query).
- **LLM reranking is more accurate but too slow.** Highest ranking quality observed, but ~52 s per query is impractical for live use.
- **LLM query rewriting hurt results.** Adding synonyms made queries less precise a useful warning for future prompt design.

Every result links back to the original opinion. Research prototype not a legal-advice tool.

Future Work

- **Scale corpus & rebalance dates** lift per-court cap (currently 2,000), extend PR coverage to 2026, add U.S. Court of Appeals for the First Circuit.
- **Listwise reranking** score K candidates in a single LLM prompt; target < 5 s/query at comparable nDCG@10.
- **Conservative query rewriting** length-conditional, additive (augment vs. replace) to reverse the observed regression.
- **Spanish-aware BM25** Snowball ES stemming and accent folding to close the diacritic gap.
- **Larger relevance pool & user study** multi-annotator labels with κ agreement, plus task-based evaluation with practicing attorneys.
- **Citation-graph signals & deploy** precomputed citation counts as a static quality prior; Docker Compose recipe for reproducible deployment.

Acknowledgements

Deepest thanks to my advisor Lisabel Rodríguez Espinosa, Ph.D., for her guidance, patience, and constructive feedback throughout this graduate project.

Thanks to the faculty of the PUPR Department of Computer Science and to the Free Law Project for maintaining the CourtListener service.

Above all, heartfelt gratitude to my family and friends none of this would have been possible without your unwavering support and patience.

References

[1] R. Gao, “Semantic Search Is Now on CourtListener,” Free Law Project, May 2026.
[2] Free Law Project, “REST API, v4.4,” CourtListener Documentation, 2026.
[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” EMNLP-IJCNLP, 2019, pp. 3982–3992.
[4] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” FnTIR, vol. 3, no. 4, pp. 333–389, 2009.
[5] G. V. Cormack et al., “Reciprocal rank fusion outperforms Condorcet and individual rank learning methods,” SIGIR, 2009, pp. 758–759.
[6] R. Nogueira and K. Cho, “Passage re-ranking with BERT,” arXiv:1901.04085, 2019.
[7] Gemma Team, Google DeepMind, “Gemma 3 Technical Report,” arXiv:2503.19786, Mar. 2025.
[8] J. R. Baron et al., “TREC-2006 Legal Track overview,” Proc. TREC 2006, 2006.
[9] O. Khattab and M. Zaharia, “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT,” SIGIR, 2020, pp. 39–48.
[10] V. Karpukhin et al., “Dense passage retrieval for open-domain question answering,” EMNLP, 2020, pp. 6769–6781.
[11] T. Pires et al., “How multilingual is multilingual BERT?” ACL, 2019, pp. 4996–5001.
[12] Sentence Transformers, “paraphrase-multilingual-MiniLM-L12-v2,” Hugging Face Model Card, 2026.
[13] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL-HLT, 2019, pp. 4171–4186.
[14] J. Johnson et al., “Billion-scale similarity search with GPUs,” IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2021.