

Retrieval for Legal AI: Complementarity and Cross-Collection Interference in Brazilian Legal Collections

Flavio Soriano, Guilherme H. G. Evangelista, Celso França,
Gisele L. Pappa, Wagner Meira Jr., Marcos André Gonçalves

Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

{flaviosoriano, guilherme.evangelista, celsofranca, glpappa, meira, mgoncalv}@dcc.ufmg.br

Abstract

Modern Legal AI systems rely on retrieval to ground answers in heterogeneous sources, including precedents, normative acts, legislative proposals, and explanatory materials. We study retrieval across five Brazilian legal collections, evaluated separately and in a pooled global index, using lexical, dense, fusion, and generative reranking architectures. The union of BM25 and dense top-10 results covers 55.4–60.3% of judged positives, 2.7–3.4 percentage points more than the better individual retriever, demonstrating substantial lexical–semantic complementarity. Reciprocal Rank Fusion (RRF) exploits part of this complementarity, but a per-query selector oracle consistently outperforms it, indicating that the preferred retrieval signal varies across queries. Pooling the collections reduces NDCG@10 by 18.5–25.9% under the original source-local judgments, showing substantial degradation when the relevant source is not known in advance; exact duplicates account for only a small fraction of this measured loss. RRF remains the strongest fixed first-stage architecture, yet we find no evidence that its gains are preserved after generative reranking at the dataset-macro level, indicating that stronger first-stage retrieval does not necessarily translate into better final rankings. These findings characterize heterogeneous legal retrieval as a multi-stage problem of source selection, signal selection, and evidence promotion.

1 Introduction

Modern legal question answering, retrieval-augmented generation, and agentic legal assistants often rely on external legal evidence for grounded reasoning (Zhong et al., 2020; Pipitone and Alami, 2024; Li et al., 2025). Retrieval therefore determines what evidence is available for attribution and citation: missing relevant evidence constrains the answer, while irrelevant or decontextualized documents may support incorrect conclusions.

Legal retrieval must balance lexical precision and semantic generalization. Exact matching remains important for statutes, case identifiers, institutional expressions, and specialized terminology, while semantically equivalent legal questions and documents may share little vocabulary (Askari et al., 2023). Sparse and dense retrievers therefore provide different signals and may recover complementary evidence, motivating hybrid architectures (França et al., 2025).

Complementarity, however, does not guarantee effective fusion. A fixed hybrid method may improve average effectiveness while failing to preserve the stronger retrieval signal for individual queries, particularly when its components disagree. Moreover, relevant evidence introduced by a stronger candidate pool may not survive downstream reranking. This creates a signal-selection problem: deciding which retrieval signal should dominate for each query and whether its gains survive subsequent ranking stages.

A second challenge arises from how legal retrieval is commonly evaluated. Collection-local benchmarks implicitly assume that the system already knows where the relevant evidence resides. Real Legal AI applications may instead search across heterogeneous repositories containing different legal genres, institutions, and levels of authority. Once these sources share an index, documents from different collections compete for the same ranking positions, potentially displacing evidence judged relevant under the source-specific task. We refer to the resulting retrieval degradation as *cross-collection interference*. This introduces a distinct *source-selection* problem: determining where relevant legal evidence should be sought when its collection is not known in advance.

We study these challenges using five Brazilian legal collections from the JUÁ benchmark (Pereira et al., 2026), covering jurisprudential precedents, administrative normative acts, legislative proposals,

and tax-related explanatory answers. We additionally pool the five corpora into a heterogeneous global index in which all 3,317 queries search the same document space, without prior knowledge of the relevant source.

Our study addresses three research questions:

RQ1. To what extent do lexical and dense retrievers provide complementary evidence, and how effectively do fixed fusion methods exploit this complementarity?

RQ2. How effective are sparse, dense, and hybrid retrieval architectures when moving from collection-local to heterogeneous global retrieval?

RQ3. To what extent are first-stage hybrid gains preserved by downstream sequence-to-sequence reranking?

We compare BM25 with six multilingual dense retrievers spanning Multilingual E5, BGE-M3, and the Qwen3-Embedding family, together with hybrid and generatively reranked architectures. Beyond aggregate effectiveness, we analyze exclusive and shared relevant evidence, component disagreement, the gap between fixed fusion and per-query selection, and retrieval behavior when all collections compete within a single global index.

Our results reveal substantial lexical–semantic complementarity, but also show that fixed fusion does not consistently preserve the stronger retrieval signal at the query level. When the collections are pooled, effectiveness drops markedly under the original source-local judgments, with exact duplicates explaining only a small fraction of the measured degradation. Finally, gains from stronger first-stage retrieval are not consistently preserved after generative reranking. Together, these results expose three coupled challenges for heterogeneous legal retrieval: selecting the source to search, the retrieval signal to trust, and the evidence to promote. Rather than proposing another retrieval architecture, our contribution is a systematic decomposition of these failure modes across heterogeneous legal sources and retrieval stages.

The main contributions of this work are:

1. a systematic evaluation of sparse, dense, hybrid, and reranked retrieval across heterogeneous Brazilian legal collections, including a global setting without prior knowledge of the relevant source;

2. a query-level analysis showing substantial lexical–semantic complementarity while exposing the limitations of fixed fusion, particularly when retrieval signals disagree;

3. an analysis of how retrieval gains propagate across stages, showing the effects of cross-collection competition and whether stronger first-stage candidate pools retain their advantage after generative reranking.

The remainder of this paper reviews related work (Section 2), presents our methodology and experimental setup (Sections 3–4), reports the results (Section 5), and discusses their implications (Section 6), followed by conclusions, limitations, and ethical considerations (Sections 7–9).

2 Related Work

Legal information retrieval requires both lexical precision and semantic generalization. BM25 remains effective for rare legal terms, citations, identifiers, and institutional expressions (Robertson and Zaragoza, 2009), while dense retrievers can recover semantically related documents with limited lexical overlap (Karpukhin et al., 2020; Reimers and Gurevych, 2019). Legal-domain and Portuguese-oriented encoders further motivate domain adaptation (Chalkidis et al., 2020; Souza et al., 2020).

Legal retrieval benchmarks span case-law, legislative, regulatory, and question-answer retrieval (Sansone and Sperli, 2022; Chalkidis et al., 2022). In Brazilian Portuguese, JUÁ provides a common interface for collections with different document types, queries, and relevance constructions (Pereira et al., 2026), including precedent retrieval in JurisTCU (Fernandes et al., 2026b), legislative retrieval in Ulysses-RFCorpus (Vitório et al., 2025), and tax-related question answering in BR-TaxQA-R (Júnior et al., 2026). Such heterogeneity motivates evaluating robustness across collections rather than on a single task.

Hybrid retrieval and reranking address different stages of the retrieval pipeline. Rank-based fusion methods such as Reciprocal Rank Fusion combine lexical and semantic rankings without requiring comparable retrieval scores; weighted variants additionally control the relative contribution of each retriever (Cormack et al., 2009). Score-level fusion instead combines retrieval scores directly, typically requiring normalization or calibration before interpolation (Bruch et al., 2024). Neural

rerankers operate at a later stage, applying more expressive relevance models to restricted candidate sets (Nogueira and Cho, 2019; Nogueira et al., 2020). These choices matter in Legal QA and Legal RAG, where retrieval determines the evidence available for grounding and attribution (Pipitone and Alami, 2024; Qi et al., 2024), and effectiveness varies across domains and query types (Thakur et al., 2021).

Document structure provides another source of retrieval evidence beyond lexical and semantic representations. Block-aware ranking has been used to assign different importance to content blocks and improve search over heterogeneous Web sites (Fernandes et al., 2007). This perspective is relevant to legal collections, whose documents differ in granularity and internal organization. Although our dense retrievers use chunks to construct document representations, the present study does not explicitly model the structural importance of sections or blocks.

Federated and vertical search have long studied retrieval over heterogeneous sources and the associated source-selection problem (Garba et al., 2023; Arguello et al., 2010). In legal retrieval, this challenge is compounded by collections and document types serving different legal functions. We connect these lines of research by studying lexical-semantic complementarity, fixed fusion, cross-collection interference, and generative reranking within a common legal retrieval setting.

3 Methodology

3.1 Legal Collections and Retrieval Regimes

We evaluate five Brazilian legal collections distributed through JUÁ: JUÁ-Juris, JurisTCU, NormasTCU, Ulysses-RFCorpus, and BR-TaxQA-R (Pereira et al., 2026). They cover judicial and administrative precedents, normative acts, legislative proposals, and tax-related explanatory material, differing substantially in corpus size, document granularity, query formulation, relevance density, and annotation procedure. Table 1 summarizes the locally prepared test collections.

JUÁ-Juris contains precedent excerpts from the *Tribunal de Contas da União* (TCU; Brazilian Federal Court of Accounts) paired with abstractive legal headnotes and binary relevance (Pereira et al., 2026); JurisTCU also contains TCU precedents but its queries include keyword searches and complete natural-language questions, with multiple graded

relevant documents (Fernandes et al., 2026b); NormasTCU shifts from precedential material to complete administrative normative acts evaluated with specialist judgments (Fernandes et al., 2026a); Ulysses-RFCorpus is a relevance-feedback corpus containing legislative proposals and information needs from members of the Brazilian Chamber of Deputies, judged by legislative consultants (Vitório et al., 2025); and BR-TaxQA-R retrieves explanatory answers to Brazilian income-tax questions (Júnior et al., 2026).

These collections therefore differ along dimensions that directly shape retrieval: legal function, document granularity, query formulation, relevance structure, and corpus scale. Queries range from short keyword expressions and abstractive headnotes to natural-language information needs, while relevance ranges from a single binary supporting document to multiple graded positives. We consequently treat the collections as distinct legal retrieval regimes rather than as interchangeable samples from a homogeneous corpus.

We evaluate two corpus configurations. In *collection-local retrieval*, each query searches only its source collection. In *global retrieval*, we merge all five corpora into a single heterogeneous index and reuse the 3,317 queries and original relevance judgments. The latter removes prior knowledge of the relevant source and forces documents with different legal functions and retrieval regimes to compete within the same ranking.

3.2 Retrieval Architectures

We compare sparse, dense, hybrid, and reranked retrieval architectures. BM25 represents sparse retrieval (Robertson and Zaragoza, 2009). Dense retrieval uses multilingual embedding models from three distinct families: Multilingual E5, BGE-M3, and Qwen3-Embedding (Wang et al., 2024; Chen et al., 2024; Zhang et al., 2025). We select these families to span different model scales and context capacities, allowing us to test whether lexical-semantic complementarity and fusion behavior persist across diverse dense retrieval systems. We evaluate both corpus-wide dense retrieval and embedding-based reranking of BM25 candidates; the latter can reorder the candidate pool but cannot recover documents absent from it.

For hybrid retrieval, we evaluate rank- and score-level fusion baselines. Standard Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) combines

Table 1: Collections used in the experiments. Positive judgments correspond to relevance scores greater than zero. Query and document lengths are medians in whitespace-separated tokens.

Collection	Documents	Queries	Positive judgments	Pos./query	Query/Doc. length	Relevance scale
JUÁ-Juris	17,147	1,714	1,714	1.00	43 / 539	Binary
JurisTCU	16,045	150	1,820	12.13	7 / 495	0-3
NormasTCU	14,466	46	299	6.50	5 / 159	0-2
Ulysses-RF	105,669	692	4,162	6.01	17 / 473	0-2
BR-TaxQA-R	715	715	1,211	1.69	18 / 84	1-2

equal-weight BM25 and Dense rankings as:

$$s_{\text{RRF}}(d) = \frac{1}{k + \text{rank}_B(d)} + \frac{1}{k + \text{rank}_D(d)}. \quad (1)$$

We also evaluate **Weighted RRF**, introducing a component balance parameter $w_B \in [0, 1]$:

$$s_{\text{WRRF}}(d) = \frac{w_B}{k + \text{rank}_B(d)} + \frac{1 - w_B}{k + \text{rank}_D(d)} \quad (2)$$

As a score-level baseline, **Normalized Score Fusion (SF)** computes a convex combination of normalized retrieval scores (Bruch et al., 2024):

$$s_{\text{SF}}(d) = \alpha \cdot \tilde{s}_B(d) + (1 - \alpha) \cdot \tilde{s}_D(d), \quad (3)$$

where $\tilde{s}(d)$ is normalized per query using min-max ($\frac{s - s_{\min}}{s_{\max} - s_{\min}}$) or z-score ($\frac{s - \mu}{\sigma}$) normalization over the top-1000 candidates; missing documents receive the minimum normalized score of that source (0 under min-max). Documents absent from a component in RRF receive no rank contribution.

We also evaluate sequence-to-sequence reranking using MonoT5 (Nogueira et al., 2020) and an mT5 checkpoint adapted on multilingual mMARCO (Xue et al., 2021; Bonifacio et al., 2021). Candidates are ranked by the log-likelihood difference:

$$s_{\text{gen}}(q, d) = \log P(y^+ | q, d) - \log P(y^- | q, d). \quad (4)$$

3.3 Lexical-Semantic Complementarity

For query q , let G_q be the judged relevant documents and B_q^K , D_q^K , and F_q^K the top- K BM25, Dense, and RRF rankings. We classify relevant documents as BM25-only, Dense-only, shared, or missed. Their union defines the relevant evidence available to either component:

$$U_q^K = G_q \cap (B_q^K \cup D_q^K), \quad R_U^K(q) = \frac{|U_q^K|}{|G_q^K|}. \quad (5)$$

The gain over the stronger component is $\Delta_{\text{union}}^K(q) = R_U^K(q) - \max\{R_B^K(q), R_D^K(q)\}$. The union is a coverage oracle, not an executable ranking.

To measure how much available evidence reaches the fused ranking, we define

$$\text{Capture}_F @O | C(q) = \frac{|F_q^O \cap U_q^C|}{|U_q^C|}. \quad (6)$$

We compute this measure for all relevant candidates and separately for BM25-only, Dense-only, and shared evidence. We also report the descriptive gap between union recall at candidate depth C and RRF recall at output depth O .

We compare fixed fusion with a retrospective per-query oracle,

$$M_{\text{oracle}}(q) = \max\{M_B(q), M_D(q)\}, \quad (7)$$

where M is primarily NDCG@10. The oracle is not deployable; it quantifies the opportunity for query-adaptive selection.

Finally, relevant-set disagreement between BM25 and Dense is one minus their Jaccard similarity:

$$\text{Disagreement}_q^K = 1 - \frac{|G_q \cap B_q^K \cap D_q^K|}{|G_q \cap (B_q^K \cup D_q^K)|}. \quad (8)$$

Rank disagreement is the median absolute BM25-Dense rank difference among relevant documents retrieved by both systems.

3.4 Cross-Collection Interference

In global retrieval, a document is *foreign* if it belongs to a collection other than the query’s source. Foreign status measures provenance rather than relevance: such documents may be related, duplicated, or useful despite not being covered by the original judgments. We quantify their prevalence by

$$\text{ForeignShare} @K(q) = \frac{|\{d \in R_q^K : c(d) \neq c(q)\}|}{K}. \quad (9)$$

We additionally measure displacement of the first relevant document,

$$\Delta r_{\text{first}}(q) = r_{\text{first}}^{\text{global}}(q) - r_{\text{first}}^{\text{local}}(q), \quad (10)$$

and count foreign documents preceding it. Finally, duplicate-aware evaluation propagates relevance to exact textual duplicates of judged relevant documents without changing the rankings. This diagnostic is related to the broader digital-library problem of identifying record replicas across repositories, which is important for content quality and repository integration (de Carvalho et al., 2006). We use a deliberately conservative exact-text criterion. Exact duplicates are identified over the prepared indexed document text, including the title: after Unicode NFC normalization, whitespace collapsing, and trimming, documents are grouped when the SHA-256 hashes of their normalized UTF-8 strings are identical.

4 Experimental Setup

4.1 Data and Retrieval Runs

Data preparation. We use the current JUÁ dataset releases publicly hosted under the *ufca-llms* organization on Hugging Face¹. All collections are converted to common corpus, query, and qrel formats, with identifier normalization, HTML and entity cleaning, whitespace normalization, and removal of documents without usable text. For analyses that require binary relevance sets, including recall, coverage, capture, and disagreement, we treat every judgment greater than zero as relevant. For NDCG, we instead preserve each collection’s original graded relevance values. The 46 Ulysses-RF queries without positive judgments remain in ranking evaluation but are excluded from measures requiring $|G_q| > 0$.

Development data. JUÁ-Juris provides the development data used for fusion parameter selection. We split its original training portion using seed 42 into 13,890 training and 1,543 local validation queries; this local validation set is created by us and is not an official JUÁ split. We use it only for fusion hyperparameter selection, while the evaluated dense checkpoints are used as distributed and are not trained on this partition.

Retrieval runs. BM25 uses $k_1 = 1.5$, $b = 0.75$, a Unicode tokenizer preserving legal identifiers, and retrieves 1,000 documents. We evaluate six multilingual dense retrievers: Multilingual E5 Large Instruct, BGE-M3, and four Qwen3-Embedding checkpoints: *qwen3-embedding-0-6b-*

jua-v2 (0.6B, fine-tuned; FT), *qwen3-embedding-4b* (4B), *qwen3-embedding-4b-jua-jurisprudencia* (4B, fine-tuned; FT), and *qwen3-embedding-8b* (8B).

All retrieve up to 1,000 documents; the fine-tuned checkpoints are used as distributed and are not retrained. Embedding reranking instead rescores BM25@1000 and returns the top 100. E5 uses 510-token content chunks with 64-token overlap, whereas BGE-M3 uses 8,190-token chunks with 256-token overlap; the Qwen3 models use 8,192-token chunks with 256-token overlap. Query and chunk embeddings are L2-normalized and scored by inner product, equivalent to cosine similarity. For documents spanning multiple chunks, the document score is the maximum similarity over its chunks,

$$S(q, d) = \max_{c \in C(d)} \hat{\mathbf{e}}_q^\top \hat{\mathbf{e}}_c.$$

This scoring rule is applied uniformly across E5, BGE-M3, and Qwen3.

Fusion parameter selection. For standard RRF, candidates from BM25@1000 and Dense@1000 are combined with equal weights, selecting $k \in \{1, 5, 10, 20, 40, 60, 80, 100, 120, 160, 200\}$ using NDCG@10 on the JUÁ-Juris validation set ($k = 10$ for 0.6B FT; $k = 1$ for E5, BGE-M3, 4B, 4B FT, and 8B). For each fusion method, hyperparameters are selected exclusively on the JUÁ-Juris validation set and transferred unchanged to all five test collections, with no collection-specific retuning. This protocol evaluates cross-regime transfer from a single development regime rather than per-collection optimization. Ties are resolved deterministically by document identifier.

Generative rerankers. MonoT5 uses *castorini/monot5-base-msmarco-10k*, and mT5 uses *unicamp-dl/mt5-base-mmarco-v2*. Both receive 1,000 candidates, use the prompt *Query: query Document: document Relevant:*, and return the top 100.

4.2 Complementarity Analysis

Complementarity is evaluated over all 3,317 test queries and six BM25–Dense pairs at $K \in \{10, 100, 1000\}$. RRF rankings contain 1,000 documents, with smaller cutoffs taken as prefixes. We report document-micro, query-macro, global query-weighted, and dataset-macro aggregations, using dataset-macro as the primary measure

¹<https://huggingface.co/ufca-llms>; dataset versions audited on August 4, 2026.

for cross-regime conclusions. Queries without positive judgments remain in ranking metrics but are excluded from recall, capture, partition, and disagreement measures.

Associations between disagreement and $\Delta_{F-\text{best}}$ use Spearman’s ρ , with 2,000 bootstrap resamples for 95% confidence intervals and 2,000 permutations for exploratory p -values; these correlations are not corrected for multiple comparisons. Runs are validated for query coverage, contiguous ranks, duplicate documents, finite scores, valid qrels, and available query text, with input/output hashes stored in manifests.

4.3 Global Retrieval

The global experiment evaluates BM25 and all six Dense–RRF pairs under collection-local and global retrieval. Global rankings are evaluated first with the original qrels and then with duplicate-aware qrels that propagate relevance to exact textual duplicates; rankings themselves remain unchanged.

We report NDCG@10, MRR@10, MAP@10, Recall@10, Recall@100, and Success@10, summarized as unweighted macro-averages over datasets. Paired effects use a dataset-stratified macro statistic, 2,000 paired-bootstrap iterations for confidence intervals, and 2,000 two-sided approximate-randomization iterations. Holm–Bonferroni correction is applied independently within each comparison family and metric. Cross-collection diagnostics use global raw rankings. First-relevant measures use a common top-100 horizon, assigning rank 101 when no relevant document appears within it.

4.4 Generative Reranking

MonoT5 and mT5 rerank either BM25@1000 or Qwen-based RRF@1000 to depth 100. The reranking experiment uses the four RRF candidate sources formed by combining BM25 with each Qwen3-Embedding retriever. Within each comparison, model, tokenizer, prompt, labels, truncation, precision, and output depth are fixed; only the candidate source changes. The primary hybrid source combines BM25 with the 4B FT dense retriever, while the three remaining Qwen-based RRF sources serve as robustness checks.

Pairwise comparisons use two-sided approximate randomization with 10,000 permutations, 95% paired-bootstrap intervals with 10,000 resamples, and Holm correction across the four Qwen-based RRF sources.

Retriever	Dense	W-RRF	SF-minmax	SF-zscore	RRF	Oracle
E5	0.3708	0.4138	0.4137	0.4137	0.4379	0.4955
BGE-M3	0.3936	0.4182	0.4137	0.4137	0.4531	0.5102
0.6B FT	0.3687	0.4513	0.4455	0.4411	0.4546	0.5110
4B	0.4475	0.4218	0.4287	0.4311	0.4723	0.5316
4B FT	0.4634	0.4364	0.4705	0.4813	0.4884	0.5459
8B	0.4577	0.4219	0.4471	0.4500	0.4828	0.5373

Table 2: Dataset-macro NDCG@10 for dense retrieval, tuned fusion baselines, RRF, and per-query component-selector oracle. BM25 achieves 0.4194. Bold indicates the strongest fixed fusion method.

4.5 Evaluation

NDCG@10 is the primary ranking metric because it accounts for rank and graded relevance. We additionally report MRR@10, MAP@10, Recall@10, Recall@100, Recall@1000, and Success@10 where applicable. Statistical tests follow the experiment-specific procedures above; disagreement correlations are treated as exploratory rather than confirmatory comparisons.

5 Results

5.1 Complementarity and Fixed Fusion

Table 2 reports dataset-macro NDCG@10 for component systems, hybrid baselines, standard RRF, and the per-query selector oracle. Standard equal-weight RRF consistently outperforms both Weighted RRF and Normalized Score Fusion across all six dense retrievers, with BM25+4B FT achieving the strongest fixed result (0.4884).

While tuning component weights on the JUÁ-Juris validation set yields high in-domain validation performance, the selected parameters (which heavily favor BM25, e.g., $w_B \geq 0.8$ or $\alpha \geq 0.7$) are strongly tailored to the JUÁ-Juris development distribution and transfer poorly to collections in which dense retrieval is substantially stronger, such as BR-TaxQA-R. Because RRF operates on ranks rather than raw scores, its transfer does not depend on score calibration across retrievers. Under the same cross-regime evaluation, standard RRF consistently outperforms the tuned weighted and score-level fusion baselines. However, the per-query selector oracle consistently exceeds RRF (reaching 0.5459 for 4B FT), indicating that fixed fusion cannot dynamically identify which retrieval signal should dominate for each query.

Performance varies substantially across collections: Dense is particularly strong on BR-TaxQA-R, whereas RRF provides clear gains on JurisTCU and Ulysses-RF. RRF therefore offers more consis-

tent cross-regime performance rather than universal dominance. Importantly, its rank-decay parameter is selected only on the JUÁ-Juris validation set and then transferred unchanged to all test collections, with no collection-specific retuning.

The complementarity analysis shows why fusion can help. At rank 10, BM25 retrieves 11.0–17.3% of judged positives missed by Dense, while Dense retrieves 10.5–15.4% missed by BM25. Their union covers 55.4–60.3% of judged positives, adding 2.7–3.4 percentage points of recall over the better individual component. At rank 1,000, union recall reaches 96.4–97.6%, indicating that the main challenge is promoting available evidence rather than finding additional deep candidates.

RRF captures 81.0–84.5% of the relevant evidence available in the component top-10 union and nearly all shared evidence, but is less effective at preserving evidence retrieved by only one component. Using candidates to depth 1,000 reinforces this result: despite union recall of 96.4–97.6%, RRF recall@10 is only 48.3–52.7%. This gap partly reflects the top-10 capacity constraint, but also shows that deep candidate availability alone does not ensure promotion.

Across all queries, RRF strictly outperforms both components in 10.3–20.3% of cases and underperforms both in only 1.8–4.4%; most often it lies between them. Its mean NDCG@10 gap to the better per-query component is consistently about -0.05 to -0.06 . Moreover, top-10 component disagreement correlates negatively with this gap ($\rho = -0.55$ to -0.44 , exploratory permutation $p \leq 0.001$): fixed equal-weight fusion is least effective precisely when BM25 and Dense disagree most strongly. This provides direct motivation for query-adaptive fusion.

5.2 Cross-Collection Interference

All retrieval configurations show lower effectiveness when the five collections are searched jointly under the original source-local relevance judgments (Table 3). Relative to collection-local retrieval, NDCG@10 decreases by 18.5–25.9% across systems, with all drops significant after Holm correction. RRF outperforms its paired dense retriever in all six configurations under global retrieval, with significant gains in five of the six comparisons; the 8B difference is not significant after Holm correction.

The diagnostics show substantial cross-collection competition, although its effects

are not uniform across architectures. Foreign documents account for 34.6–45.9% of the top 10 and 45.1–54.9% of the top 100. No system dominates all diagnostics: 4B Dense has the lowest foreign-document share at both cutoffs and the smallest first-relevant displacement, whereas 4B FT RRF achieves the highest NDCG@10 and the fewest foreign documents before the first relevant result. Moreover, the paired Dense retriever exhibits a smaller first-relevant displacement than RRF in all six configurations. Thus, RRF’s higher ranking effectiveness does not imply uniformly lower provenance intrusion or relevant-document displacement.

Duplicate-aware evaluation raises global NDCG@10 by only 0.003–0.009, despite these gains being statistically significant, recovering only a small fraction of the measured local-to-global loss. Exact duplicates therefore account for little of this measured degradation. The remaining loss may reflect competition among heterogeneous documents, unjudged cross-collection relevance, near-duplicate or semantically equivalent content, or mismatch between topical similarity and the legal source expected by the original task.

5.3 Generative Reranking

Finally, we test whether the first-stage gains from RRF are preserved after generative reranking. We find no evidence that these gains translate into improved dataset-macro effectiveness after reranking. MonoT5 reaches dataset-macro NDCG@10 of 0.3252 with BM25 candidates and 0.3240 with RRF 4B FT candidates. mT5 is stronger overall, reaching 0.3915 with BM25 candidates and 0.3868 with RRF candidates. Thus, replacing BM25 with the stronger hybrid candidate pool changes dataset-macro NDCG@10 by only -0.0012 for MonoT5 and -0.0047 for mT5. Neither difference is significant at the dataset-macro level after Holm correction.

For the primary 4B FT RRF contrast, significant decreases occur on JurisTCU and Ulysses-RF, but these collection-specific effects do not establish a significant degradation at the dataset-macro level. Overall, stronger hybrid candidate coverage does not automatically translate into stronger top-ranked results after sequence-to-sequence reranking.

Table 3: Dataset-macro NDCG@10 and cross-collection diagnostics. F@10 and F@100 denote foreign-document share, and F-before-rel. denotes the number of foreign documents preceding the first relevant result. Higher NDCG and lower diagnostic values are better.

Retriever	System	NDCG@10			Global diagnostics			
		Local	Global	Dup.-aware	F@10	F@100	First-rel. Δ	F-before-rel.
–	BM25	0.4194	0.3106	0.3166	0.4182	0.5179	10.8146	6.4739
E5	Dense	0.3708	0.2943	0.2973	0.3931	0.4882	7.2655	5.9704
	RRF	0.4379	0.3363	0.3418	0.3909	0.4904	7.9292	5.8440
BGE-M3	Dense	0.3936	0.3187	0.3266	0.3839	0.4973	7.1137	5.4845
	RRF	0.4531	0.3495	0.3568	0.3854	0.4942	7.8851	6.1071
0.6B FT	Dense	0.3687	0.2742	0.2832	0.4594	0.5489	8.8662	6.2716
	RRF	0.4546	0.3487	0.3567	0.4170	0.5172	9.3379	5.0065
4B	Dense	0.4475	0.3647	0.3721	0.3463	0.4507	6.0319	5.0006
	RRF	0.4723	0.3755	0.3823	0.3677	0.4711	6.9675	5.0039
4B FT	Dense	0.4634	0.3727	0.3797	0.3707	0.4778	6.7665	4.8461
	RRF	0.4884	0.3876	0.3946	0.3787	0.4810	7.2883	4.6383
8B	Dense	0.4577	0.3702	0.3777	0.3887	0.4805	7.5202	5.9732
	RRF	0.4828	0.3789	0.3860	0.3899	0.4873	8.1093	5.1684

6 Discussion

RQ1: Lexical–semantic complementarity and fusion. Lexical and dense retrievers recover complementary relevant evidence, yet fixed fusion exploits this complementarity only partially. Their top-10 union consistently increases coverage over the better individual component, while the per-query selector oracle remains above RRF across all dense retrievers. Moreover, RRF is least effective relative to the stronger component when the two rankings disagree most. Thus, complementarity creates an opportunity for hybrid retrieval, but fixed equal-weight fusion cannot consistently determine which retrieval signal should dominate for each query.

RQ2: Collection-local versus global retrieval. Moving from collection-local to global retrieval reveals a distinct source-selection problem when the relevant collection is not known in advance. The substantial degradation under global retrieval shows that collection-local evaluation masks this decision problem by removing competition among heterogeneous document types, institutions, and legal functions. This degradation persists across all six Dense–RRF configurations and multiple dense model families, suggesting that the local-to-global effect is not specific to a single retriever. Importantly, these diagnostics characterize provenance competition, not the relevance of foreign documents: the original judgments cannot determine whether unjudged cross-collection results constitute useful evidence.

RRF remains the strongest architecture in NDCG@10 under global retrieval, but the diagnostic results reveal a more nuanced picture. Dense retrieval shows smaller first-relevant displacement across all paired configurations, whereas RRF achieves the strongest overall ranking effectiveness. Cross-collection retrieval behavior is therefore multidimensional: stronger ranking effectiveness does not necessarily imply less provenance intrusion or smaller displacement of source-local relevant documents.

RQ3: Preservation of first-stage gains after reranking. Stronger first-stage candidate pools do not necessarily yield better final rankings. Neither evaluated reranker shows evidence of preserving the RRF advantage at the dataset-macro level, suggesting that gains from complementary first-stage retrieval are not reliably propagated through reranking. Domain mismatch, binary relevance modeling, or loss of lexical and provenance signals are plausible explanations, but our experiments do not isolate their causal effects.

Architectural implications for Legal AI. These findings suggest that Legal QA and RAG pipelines should treat retrieval as an orchestrated sequence of decisions rather than as a single ranking operation. When multiple heterogeneous repositories are available, source routing or metadata-based filtering can reduce unnecessary competition before ranking. Within the selected search space, lexical and dense signals may need to be combined adaptively, since a fixed fusion policy does not con-

sistently preserve the stronger signal across queries. Finally, downstream rerankers should be evaluated not only by their standalone ranking effectiveness, but also by whether they preserve complementary evidence and provenance information made available by the first-stage retriever. These results suggest that robust retrieval for Legal AI may benefit from coordinating source selection, signal selection, and evidence promotion across stages.

7 Conclusion

We studied retrieval for Legal AI across five heterogeneous Brazilian legal collections and a global setting in which the relevant source is not known in advance, comparing sparse, dense, hybrid, and reranked retrieval architectures. Our results expose three coupled challenges in heterogeneous legal retrieval. First, lexical and dense retrievers provide complementary evidence, but fixed RRF cannot consistently determine which signal should dominate for each query, motivating query-adaptive fusion. Second, pooling heterogeneous sources into a common index produces substantial effectiveness degradation under the original source-local judgments, while provenance diagnostics reveal different patterns of cross-collection competition across architectures. This motivates collection routing and provenance-aware retrieval when the relevant source is not known in advance.

Generative reranking further shows that stronger candidate evidence is not necessarily preserved downstream. Taken together, these findings characterize robust Legal AI retrieval as a multi-stage evidence-selection problem involving three coupled decisions: where to search, which retrieval signals to trust, and which retrieved evidence to promote. Future work should investigate methods for source selection, signal selection, and evidence promotion in heterogeneous Legal AI.

8 Limitations

Our model comparisons cover multiple retrieval architectures, but do not isolate the contribution of individual design choices. Across Multilingual E5, BGE-M3, and Qwen3-Embedding, differences in training, context length, chunking, query formatting, and model scale vary jointly; the comparison therefore concerns complete retrieval systems rather than the isolated effect of encoder architecture. The downstream reranking stage has a similarly restricted scope: we evaluate only

two T5-family models, without fine-tuning on the evaluated legal judgments and with 512-token joint inputs, and do not consider legal cross-encoders, passage-level retrieval, or evidence-aware truncation. Moreover, all collections are Brazilian Portuguese datasets distributed through JUÁ, limiting generalization to other jurisdictions, languages, and legal information environments.

Original collection-specific judgments may omit useful cross-collection evidence, and collection origin indicates provenance rather than relevance. Consequently, the measured local-to-global degradation should not be interpreted as an estimate of the amount of genuinely irrelevant cross-collection retrieval without additional cross-collection relevance judgments. Our duplicate-aware evaluation provides a sensitivity analysis for one source of incomplete judgments, showing that exact duplicates explain only a small fraction of the measured degradation. However, it captures only exact matches and therefore does not account for near-duplicates, revisions, semantically equivalent documents, or genuinely relevant evidence from other collections. Collection-local retrieval also represents an idealized setting with perfect source selection.

The relevance judgments used in our experiments operationalize benchmark-specific relevance rather than a complete notion of legal relevance in practice, which may also depend on authority, validity, temporal applicability, jurisdiction, and procedural context. Our effectiveness metrics should therefore be interpreted as agreement with the available judgments, not as a complete assessment of the legal appropriateness of retrieved evidence. Likewise, the retrieval signals studied here are not substitutes for legal interpretation: dense similarity captures representational proximity, while fusion combines retrieval evidence without resolving conflicts between legal authorities or determining which source should prevail. Max-over-chunks scoring may also promote a document based on a highly similar local passage without representing its full legal context. Finally, our evaluation concerns document retrieval rather than downstream answer correctness, groundedness, citation validity or the ability of a Legal QA or RAG system to use the retrieved evidence appropriately.

9 Ethical Considerations

Legal retrieval operates in a high-stakes domain. Retrieved documents may be outdated, biased,

superseded, or inappropriate for the relevant jurisdiction or legal context; benchmark relevance does not imply legal validity, authority, or applicability. The evaluated systems should therefore be treated as research or decision-support tools rather than substitutes for qualified legal professionals.

Cross-collection retrieval further increases the importance of provenance: primary authorities, legislative proposals, administrative precedents, and explanatory materials may be topically similar while serving different legal functions. Practical systems should expose source, document type, institution, date, and legal status, support human verification, and preserve attribution when equivalent or duplicated documents occur across repositories.

Finally, collections and models may reproduce biases from source availability, institutional curation, annotation, and pretraining data. Responsible deployment requires monitoring both systematic omissions and incorrect retrievals.

Acknowledgments

This work was supported by CNPq, CAPES, INCT-TILDIAR (grant no. 408490/2024-1), CIIA-Saude (process No. PPE-00030-21), UNIMED-BH, FAPEMIG, AWS, NVIDIA, Google and FAPESP.

References

- Jaime Arguello, Fernando Diaz, and Jean-François Paiement. 2010. [Vertical selection in the presence of unlabeled verticals](#). In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*, pages 691–698, New York, NY, USA. Association for Computing Machinery.
- Arian Askari, Amin Abolghasemi, Gabriella Pasi, Wessel Kraaij, and Suzan Verberne. 2023. [Injecting the BM25 score as text improves BERT-based re-rankers](#). In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, pages 66–83, Berlin, Heidelberg. Springer-Verlag.
- Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. [mmarco: A multilingual version of ms marco passage ranking dataset](#). *Preprint*, arXiv:2108.13897.
- Sebastian Bruch, Siyu Gai, and Amir Ingber. 2024. [An analysis of fusion functions for hybrid retrieval](#). *ACM Trans. Inf. Syst.*, 42(1).
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. [Legal-bert: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A benchmark dataset for legal language understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.
- Moisés G. de Carvalho, Marcos André Gonçalves, Alberto H. F. Laender, and Altigran S. da Silva. 2006. [Learning to deduplicate](#). In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 41–50. IEEE.
- Leandro Carísio Fernandes, Marcus Vinícius Borela de Castro, Leandro dos Santos Ribeiro, Leonardo Augusto da Silva Pacheco, and Edans Flávio de Oliveira Sandes. 2026a. [NormasTCU — a brazilian portuguese ir dataset and an evaluation of llm-as-a-judge for relevance assessment](#). *Preprint*, arXiv:2608.27746.
- Leandro Carísio Fernandes, Leandro dos Santos Ribeiro, Marcus Vinícius Borela de Castro, Leonardo Augusto da Silva Pacheco, and Edans Flávio de Oliveira Sandes. 2026b. [Juristcu: a brazilian portuguese information retrieval dataset with query relevance judgments](#). *Language Resources and Evaluation*, 60(1):23.
- Celso França, Gestefane Rabbi, Thiago Salles, Washington Cunha, Leonardo Rocha, and Marcos André Gonçalves. 2025. [Optimizing tail-head trade-off for extreme multi-label text classification \(XMTC\) with RAG-labels and a dynamic two-stage retrieval and fusion pipeline](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, pages 1392–1401, New York, NY, USA. Association for Computing Machinery.
- Adamu Garba, Shengli Wu, and Shah Khalid. 2023. [Federated search techniques: an overview of the trends and state of the art](#). *Knowl. Inf. Syst.*, 65(12):5065–5095.

- Juvenal Domingos Júnior, Augusto Faria, E. Seiti de Oliveira, Erick de Brito, Matheus Teotonio, Andre Assumpção, Diedre Carmo, Roberto Lotufo, and Jayr Pereira. 2026. [Br-taxqa-r: A dataset for question answering with references for brazilian personal income tax law, including case law](#). In *Intelligent Systems: 35th Brazilian Conference, BRACIS 2025, Fortaleza-CE, Brazil, September 29 – October 2, 2025, Proceedings, Part I*, pages 208–222, Berlin, Heidelberg. Springer-Verlag.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.
- Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025. [LegalAgentBench: Evaluating LLM agents in legal domain](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria. Association for Computational Linguistics.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with bert](#). *Preprint*, arXiv:1901.04085.
- Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. [Document ranking with a pre-trained sequence-to-sequence model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.
- Jayr Pereira, Leandro Fernandes, Erick de Brito, Roberto Lotufo, and Luiz Bonifacio. 2026. [Juá – a benchmark for information retrieval in brazilian legal text collections](#). *Preprint*, arXiv:2604.06098.
- Nicholas Pipitone and Ghita Houir Alami. 2024. [Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain](#). *Preprint*, arXiv:2408.10343.
- Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. [Model internals-based answer attribution for trustworthy retrieval-augmented generation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Carlo Sansone and Giancarlo Sperli. 2022. [Legal information retrieval systems: State-of-the-art and open issues](#). *Information Systems*, 106:101967.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. [Bertimbau: Pretrained bert models for brazilian portuguese](#). In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*, pages 403–417. Springer.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
- Douglas Vitório, Ellen Souza, Lucas Martins, Nádia F. F. da Silva, André Carlos Ponce de Leon de Carvalho, Adriano L. I. Oliveira, and Francisco Edmundo de Andrade. 2025. [Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the Brazilian Chamber of Deputies](#). *Language Resources and Evaluation*, 59(2):1257–1277.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual e5 text embeddings: A technical report](#). *arXiv preprint arXiv:2402.05672*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. [JEC-QA: A legal-domain question answering dataset](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.