Fusão Híbrida para Recuperação de Informação Jurídica Brasileira

Flávio Soriano, Guilherme Evangelista, Celso França, Gisele Pappa, Wagner Meira Jr. e Marcos Gonçalves

[1]Universidade Federal de Minas Gerais, Brasil

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

Abstract. *Brazilian legal retrieval involves heterogeneous search regimes that simultaneously require precise lexical matching and semantic generalization. While hybrid retrieval has become common in general information retrieval, it remains unclear when lexical and semantic signals effectively complement each other in diverse Brazilian legal collections. We investigate this question on JUÁ, a benchmark covering multiple legal retrieval scenarios, including jurisprudence, normative acts, legislative retrieval, and tax-related question answering. We compare BM25, Qwen3 dense retrievers, legal fine-tuned variants, rerankers, and hybrid fusion via Reciprocal Rank Fusion (RRF). Our hypothesis is that lexical-dense fusion can provide greater robustness across heterogeneous legal collections by combining complementary retrieval signals without requiring direct score calibration. Results show that RRF-based systems achieve the best aggregate effectiveness: a legal fine-tuned 4B model leads NDCG@10 and MRR@10, while an 8B model leads MAP@10 and Recall@10. Paired statistical tests show that RRF yields robust gains over BM25 and part of the corresponding dense baselines, although not universally dominating reranking in every individual dataset. Overall, the results suggest that retrieval effectiveness in the legal domain depends strongly on the underlying retrieval regime, and that hybrid lexical-semantic fusion provides a particularly robust configuration for heterogeneous legal search.*

Resumo. *A recuperação de informação jurídica brasileira envolve regimes heterogêneos de busca que exigem simultaneamente correspondência lexical precisa e generalização semântica. Embora a recuperação híbrida seja consolidada na recuperação de informação geral, ainda não está claro quando sinais lexicais e semânticos se complementam efetivamente em coleções jurídicas brasileiras diversas. Investigamos essa questão no benchmark JUÁ, que abrange múltiplos cenários: jurisprudência, atos normativos, recuperação legislativa e perguntas tributárias. Comparamos BM25, recuperadores densos da família Qwen3, variantes ajustadas ao domínio jurídico, reranqueadores e fusão híbrida via Reciprocal Rank Fusion (RRF). Nossa hipótese é que a fusão léxico-densa oferece maior robustez ao combinar sinais complementares sem necessidade de calibração direta de scores. Os resultados mostram que sistemas baseados em RRF alcançam a melhor efetividade agregada: o modelo jurídico ajustado de 4B lidera em NDCG@10 e MRR@10, enquanto o de 8B lidera em MAP@10 e Recall@10. Testes estatísticos pareados indicam*

ganhos robustos do RRF sobre BM25 e parte dos modelos densos, embora não domine todos os reranqueadores individualmente. No geral, a fusão híbrida léxico-semântica revela-se uma configuração particularmente robusta para busca jurídica heterogênea.

1. Introdução

A recuperação de informação jurídica em português brasileiro envolve cenários heterogêneos, que vão da busca jurisprudencial e normativa à recuperação legislativa e perguntas tributárias. Nesses contextos, a relevância de um documento frequentemente depende de correspondência lexical precisa, como termos técnicos, citações normativas e nomes de órgãos, ao mesmo tempo em que consultas em linguagem natural e paráfrases exigem modelos capazes de capturar relações semânticas. Essa tensão torna o domínio jurídico particularmente adequado para investigar a interação entre evidências lexicais, esparsas, e semânticas, densas, em coleções com diferentes regimes de recuperação e necessidades de informação.

Embora estratégias híbridas tenham se consolidado na recuperação de informação geral, ainda há pouca evidência sistemática sobre como sinais lexicais e semânticos se complementam em coleções jurídicas brasileiras heterogêneas. Diferentemente de avaliações centradas em uma única coleção, tarefa ou regime jurídico, este trabalho investiga se uma estratégia simples de fusão por RRF mantém robustez entre regimes jurídicos estruturalmente distintos. Essa questão permanece empiricamente aberta porque a literatura mostra que a eficácia de métodos neurais e híbridos varia conforme domínio, tarefa e disponibilidade de modelos especializados [Thakur et al. 2021, Chen et al. 2022, Louis et al. 2025]. No caso do JUÁ [Pereira et al. 2026], essa questão é particularmente relevante porque os subconjuntos variam não apenas em domínio, mas também em tipo documental, estilo de consulta, granularidade dos julgamentos de relevância e grau de dependência entre correspondência lexical e semântica.

Neste trabalho, investigamos essa questão no benchmark JUÁ [Pereira et al. 2026], uma suíte que reúne diferentes regimes de recuperação jurídica brasileira. Nossa pergunta de pesquisa central é: *em cenários jurídicos heterogêneos, a combinação explícita entre evidência lexical e semântica produz sistemas de recuperação mais robustos do que abordagens puramente densas ou estratégias de reranqueamento?* Para respondê-la, avaliamos BM25, recuperadores densos da família Qwen3-Embedding [Zhang et al. 2025], variantes ajustadas ao domínio jurídico, *pipelines* de reranqueamento e fusão híbrida via *Reciprocal Rank Fusion* (RRF). Nossa hipótese é que a fusão léxico-densa por RRF, ao combinar rankings por posição sem exigir calibração direta de *scores*, pode aproveitar complementaridades entre evidência lexical e semântica, produzindo maior robustez agregada entre coleções jurídicas distintas, mesmo quando não domina todos os subconjuntos individualmente.[1]

Os resultados sustentam essa hipótese no cenário agregado. RRF 4B FT alcança os maiores valores de *NDCG@10* e *MRR@10*, enquanto RRF 8B lidera em *MAP@10* e *Recall@10*. Os testes estatísticos pareados indicam ganhos robustos sobre BM25 e ganhos sobre parte dos recuperadores densos correspondentes, embora a superioridade sobre Rerank não seja uniforme em todos os regimes. Assim, os resultados apontam RRF não

[1] https://github.com/flaviosoriano/fusao-hibrida-recuperacao-informacao-juridica-brasileira

como uma solução universalmente dominante, mas como uma estratégia simples e estável para combinar sinais lexicais e semânticos em um benchmark jurídico heterogêneo.

As principais contribuições deste artigo são:

1. investiga sistematicamente a complementaridade entre sinais lexicais e semânticos em múltiplos regimes de recuperação jurídica brasileira;
2. apresenta uma avaliação comparativa entre recuperação esparsa, densa, reranqueamento e fusão híbrida no benchmark JUÁ;
3. analisa empiricamente em quais cenários evidências lexicais e semânticas atuam de forma complementar;
4. mostra que estratégias híbridas via RRF produzem maior robustez agregada entre coleções heterogêneas, mesmo sem dominância universal em todos os subconjuntos.

2. Trabalhos Relacionados

A recuperação de informação jurídica envolve equilibrar correspondência terminológica precisa com compreensão semântica de consultas em linguagem natural. Métodos lexicais como BM25 [Robertson and Zaragoza 2009] permanecem fortes no domínio jurídico por capturarem termos técnicos, citações normativas e jargões institucionais, enquanto modelos densos, como DPR [Karpukhin et al. 2020] e Sentence-BERT [Reimers and Gurevych 2019], ampliaram a recuperação semântica em cenários de baixa sobreposição lexical. Essa complementaridade motivou adaptações de modelos para *corpora* jurídicos, como LEGAL-BERT [Chalkidis et al. 2020] e abordagens baseadas em BERTimbau [Souza et al. 2020].

Para combinar sinais lexicais e semânticos, a literatura recente explora estratégias de fusão e reranqueamento. O *Reciprocal Rank Fusion* (RRF) [Cormack et al. 2009] combina rankings sem exigir calibração direta de *scores*, enquanto o reranqueamento neural [Nogueira and Cho 2019] reaplica modelos neurais sobre candidatos inicialmente recuperados. Como mostrado em *benchmarks* heterogêneos como BEIR [Thakur et al. 2021], a eficácia varia conforme domínio e tipo de consulta. O benchmark JUÁ [Pereira et al. 2026] formalizou essa diversidade no cenário jurídico brasileiro. Nosso trabalho investiga empiricamente quando a fusão via RRF produz maior robustez do que recuperação densa isolada ou reranqueamento no direito brasileiro.

3. Metodologia

3.1. Preparação dos Dados e Benchmark

Este trabalho avalia métodos de recuperação de informação no benchmark JUÁ, uma suíte pública para recuperação jurídica em português brasileiro [Pereira et al. 2026]. O benchmark reúne coleções com diferentes tipos de documentos e necessidades de informação, incluindo jurisprudência, atos normativos, documentos legislativos e perguntas tributárias. Essa heterogeneidade torna o JUÁ adequado para analisar não apenas o desempenho médio dos sistemas, mas também como diferentes mecanismos de recuperação se comportam em regimes jurídicos distintos.

A tarefa segue a formulação clássica de recuperação ad hoc. Dado um conjunto de consultas Q, um corpus documental D e julgamentos de relevância $\mathcal{R}$, cada sistema deve produzir, para cada consulta, uma lista ordenada de documentos. A qualidade do

ranking é avaliada pela comparação entre os documentos recuperados e os julgamentos de relevância disponíveis. Todos os métodos avaliados neste trabalho utilizam a mesma preparação local dos dados, de modo que as diferenças observadas reflitam os mecanismos de recuperação, e não variações de pré-processamento ou de protocolo experimental.

3.1.1. Coleções avaliadas

A avaliação considera cinco subconjuntos associados ao JUÁ: JUÁ-Juris, juris_tcu, normas_tcu, ulysses_rf e br_taxqa. Os dois primeiros são voltados à recuperação de jurisprudência, enquanto normas_tcu cobre atos normativos do TCU, ulysses_rf representa recuperação legislativa e br_taxqa envolve perguntas em linguagem natural no domínio tributário. Assim, os datasets diferem quanto ao tipo documental, ao formato das consultas e à estrutura dos julgamentos de relevância, permitindo avaliar a robustez dos métodos em cenários jurídicos variados.

Todos os subconjuntos foram convertidos para um formato comum de recuperação, composto por arquivos *corpus.jsonl*, *queries.jsonl* e *qrels_*.tsv*. A preparação local inclui normalização de identificadores, remoção de artefatos simples de HTML, decodificação de entidades, normalização de espaços e conversão dos julgamentos para um formato compatível com o avaliador. Documentos ou consultas sem conteúdo textual útil foram descartados.

3.1.2. Características dos dados

A Tabela 1 resume os subconjuntos avaliados, incluindo tamanho do corpus, número de consultas, quantidade de julgamentos de relevância no conjunto de teste e escala de relevância utilizada. As coleções diferem quanto ao tipo documental, formato das consultas e estrutura dos julgamentos de relevância. Em particular, BR-TaxQA-R enfatiza consultas em linguagem natural, enquanto JurisTCU, NormasTCU e Ulysses-RFCorpus possuem escalas graduadas de relevância.

Tabela 1. Características dos dados preparados localmente para avaliação.

Dataset	Documentos	Consultas	Qrels teste	Escala
JUÁ-Juris	17147	17147	1714	1
JurisTCU	16045	150	2250	0–3
NormasTCU	14469	46	812	0–2
Ulysses-RFCorpus	105669	692	8205	0–2
BR-TaxQA-R	715	715	1211	1–2

3.1.3. Métricas

A avaliação utiliza métricas de ranking no corte 10: *NDCG@10*, *MRR@10*, *MAP@10* e *Recall@10*. *NDCG@10* mede a qualidade posicional considerando graus de relevância; *MRR@10* enfatiza a posição do primeiro documento relevante; *MAP@10* avalia a precisão acumulada nos pontos em que documentos relevantes são recuperados; e *Recall@10* mede a cobertura de documentos relevantes entre os dez primeiros resultados.

Essas métricas permitem distinguir ganhos de ordenação no topo do ranking de ganhos de cobertura, fornecendo uma base adequada para comparar BM25, modelos densos, reranqueamento e fusão híbrida [Kamalloo et al. 2024, Thakur et al. 2021].

3.2. Métodos

Este trabalho propõe e avalia uma estratégia de fusão híbrida para recuperação de informação jurídica brasileira baseada em *Reciprocal Rank Fusion* (RRF). O objetivo é investigar, em um benchmark jurídico heterogêneo, se a combinação entre evidência léxico-sintática e evidência semântica produz rankings mais robustos do que métodos isolados ou pipelines neurais de reranqueamento. A proposta parte da hipótese de que a recuperação jurídica se beneficia da combinação entre sensibilidade à forma textual dos documentos e capacidade de generalização semântica, motivando a combinação de métodos lexicais, como BM25, com modelos densos de recuperação [Nigam et al. 2022].

Dessa forma, organizamos os métodos avaliados em torno de quatro famílias: recuperação lexical, recuperação densa, reranqueamento e fusão por RRF. A proposta central do artigo é avaliar a fusão léxico-densa como uma alternativa simples e robusta para combinar sinais lexicais e semânticos em recuperação jurídica brasileira.

3.2.1. Recuperação lexical e recuperação densa

A recuperação lexical é representada por BM25, utilizado tanto como baseline quanto como uma das entradas da fusão híbrida. BM25 atribui pontuação aos documentos a partir da correspondência entre os termos da consulta e os termos do documento, ponderando frequência lexical, frequência inversa no corpus e normalização por comprimento.

A recuperação densa representa consultas e documentos em um espaço vetorial compartilhado. Cada consulta e cada documento são codificados como vetores, e a ordenação é produzida a partir de uma medida de similaridade entre essas representações. Esse mecanismo busca capturar proximidade semântica mesmo quando há baixa sobreposição lexical.

Neste trabalho, avaliamos dois recuperadores densos gerais, Qwen3-Embedding-4B e Qwen3-Embedding-8B, e duas variantes adaptadas ao domínio jurídico, Qwen3-Embedding-4B FT e Qwen3-Embedding-0.6B FT. No restante do artigo, referimo-nos a esses modelos, respectivamente, como 4B, 8B, 4B FT e 0.6B FT. As variantes FT foram disponibilizadas no ecossistema JUÁ e derivam de ajuste fino de modelos Qwen para recuperação jurídica brasileira, conforme descrito no paper do benchmark [Pereira et al. 2026]. Neste artigo, a adaptação ao domínio não é tratada como substituta da recuperação lexical, mas como uma forma de fortalecer o componente semântico. Assim, uma das questões avaliadas é se a especialização jurídica do modelo denso reduz ou preserva a necessidade de combiná-lo com evidência lexical.

3.2.2. Fusão híbrida por RRF

O componente central da proposta é a fusão de rankings por RRF. A motivação é combinar listas de candidatos produzidas por mecanismos distintos, preservando documentos recuperados por evidência lexical ou semântica. Em um benchmark jurídico heterogêneo,

essa propriedade é desejável porque diferentes coleções podem favorecer diferentes fontes de evidência. Utilizamos RRF porque a fusão é feita com base nas posições dos documentos nos rankings de entrada, e não na combinação direta dos *scores*. Isso é importante porque *scores* de BM25 e similaridades vetoriais não estão naturalmente calibrados na mesma escala. Para um documento d, a pontuação por RRF é definida como $\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)}$, em que R é o conjunto de rankings combinados, $\mathrm{rank}_r(d)$ é a posição do documento d no ranking r, e k é uma constante que controla o peso relativo das primeiras posições. Documentos bem posicionados em múltiplos rankings recebem maior pontuação, mas documentos recuperados por apenas um dos mecanismos também podem ser preservados no ranking final.

Em cada variante de fusão, combinamos um ranking lexical produzido por BM25 com um ranking denso produzido por um modelo de embeddings $m \in \mathcal{M}$. Assim, a família de sistemas RRF avaliada neste trabalho pode ser representada por $\mathrm{RRF}_m = \mathrm{RRF}(r_{\mathrm{BM25}}, r_m)$, em que r_{BM25} é o ranking produzido por BM25 e r_m é o ranking produzido pelo recuperador denso m. Essa formulação torna explícito que a proposta não depende de um único modelo denso específico: ela permite avaliar a mesma estratégia de fusão entre evidência léxico-sintática e evidência semântica com diferentes recuperadores densos, incluindo modelos gerais e variantes adaptadas ao domínio jurídico.

3.2.3. Sistemas comparados no benchmark

Para avaliar a proposta, comparamos sistemas que isolam diferentes mecanismos de recuperação: BM25, recuperação densa, recuperação densa adaptada ao domínio, reranqueamento sobre candidatos lexicais e fusão híbrida por RRF. Seja $\mathcal{M}$ o conjunto de modelos densos avaliados: $\mathcal{M} = \{\text{4B}, \text{8B}, \text{4B FT}, \text{0.6B FT}\}$.

Para cada $m \in \mathcal{M}$, avaliamos três usos possíveis do modelo: **Dense** m, isto é, recuperação densa direta; **Rerank** m, isto é, reranqueamento de candidatos recuperados por BM25; e **RRF** m, isto é, fusão entre BM25 e o ranking denso produzido por m. A Tabela 2 resume essas famílias de sistemas.

Tabela 2. Famílias de sistemas avaliadas.

Família	Forma	Mecanismo avaliado
Esparso	BM25	Correspondência léxico-sintática.
Denso	Dense m, para $m \in \mathcal{M}$	Correspondência semântica por embeddings.
Rerank	Rerank m, para $m \in \mathcal{M}$	Reordenação neural de candidatos recuperados por BM25.
Fusão RRF	RRF m, para $m \in \mathcal{M}$	Fusão entre evidência léxico-sintática e semântica.

4. Experimentos

4.1. Configuração Experimental

Os sistemas definidos na Seção 3.2.3 foram executados nos cinco corpora avaliados, usando a preparação local dos dados descrita na Seção 3.1.1. A seguir, descrevemos os parâmetros de execução que diferenciam BM25, Dense, Rerank e RRF.

BM25 foi configurado com $k_1 = 1{,}5$, $b = 0{,}75$ e retorno dos 1000 documentos mais bem classificados por consulta. A tokenização utiliza uma expressão regular

Unicode que preserva números e sequências com hífen, ponto ou barra, o que é relevante para citações normativas e identificadores jurídicos. A recuperação densa utiliza modelos da família Qwen3-Embedding [Zhang et al. 2025], carregados com SentenceTransformer [Wolf et al. 2020], retornando os 1000 documentos mais próximos por consulta.

A fusão RRF combina sempre dois rankings: BM25 e um modelo denso. Utilizamos RRF simples, sem pesos diferenciados, com $k = 60$ e peso $1{,}0$ para cada entrada. O valor $k = 60$ segue a recomendação original de Cormack et al. [Cormack et al. 2009], que o apresentam como uma escolha padrão robusta para RRF. Essa escolha evita introduzir hiperparâmetros adicionais dependentes de dataset e mantém a comparação focada na complementaridade entre o sinal lexical e o sinal semântico, em vez de em uma etapa de calibração ou otimização dos pesos da fusão. A saída contém os 100 documentos mais bem classificados por consulta. Documentos ausentes em um dos rankings não contribuem naquele ranking, e empates são resolvidos pelo identificador do documento.

O reranqueamento, usado como baseline comparativo, é baseado em embeddings bi-encoder com os mesmos modelos de $\mathcal{M}$: para cada consulta, recebe os 1000 candidatos recuperados por BM25, substitui o score original pelo score do reranqueador e retorna os 100 primeiros documentos, usando o rank BM25 original e o identificador do documento como critérios de desempate.

4.2. Resultados Experimentais

Esta seção apresenta os resultados dos sistemas avaliados nos cinco subconjuntos do benchmark e no agregado geral. Primeiro, reportamos o desempenho agregado de todos os sistemas. Em seguida, apresentamos os testes estatísticos pareados em *NDCG@10*. Por fim, resumimos quais sistemas obtiveram os melhores valores por dataset e métrica. As métricas reportadas são *NDCG@10*, *MRR@10*, *MAP@10* e *Recall@10*.

4.2.1. Resultado agregado

Tabela 3. Resultados agregados dos sistemas avaliados. Todas as métricas são calculadas em @10.

Família	Sistema	NDCG	MRR	MAP	Recall
Esparso	BM25	0.4137	0.5223	0.3000	0.4416
Denso	Dense 4B	0.4445	0.5474	0.3329	0.4729
Denso	Dense 8B	0.4507	0.5557	0.3399	0.4722
Denso	Dense 4B FT	0.4562	0.5676	0.3365	0.4845
Denso	Dense 0.6B FT	0.3647	0.4670	0.2591	0.4118
Rerank	Rerank 4B	0.4480	0.5503	0.3352	0.4779
Rerank	Rerank 8B	0.4545	0.5565	0.3431	0.4784
Rerank	Rerank 4B FT	0.4623	0.5740	0.3413	0.4909
Rerank	Rerank 0.6B FT	0.3755	0.4804	0.2678	0.4220
Fusão	RRF 4B	0.4594	0.5701	0.3431	0.4830
Fusão	RRF 8B	0.4690	0.5650	**0.3513**	**0.5009**
Fusão	RRF 4B FT	**0.4744**	**0.5835**	0.3508	0.5006
Fusão	RRF 0.6B FT	0.4326	0.5363	0.3131	0.4708

Os resultados agregados mostram que as melhores configurações observadas pertencem à família de fusão por RRF. RRF 4B FT obtém os maiores valores de *NDCG@10* e *MRR@10*, com 0,4744 e 0,5835, respectivamente. RRF 8B alcança os maiores valores de *MAP@10* e *Recall@10*, com 0,3513 e 0,5009.

BM25 apresenta desempenho agregado inferior aos modelos densos, reranqueadores e fusões. Os modelos densos gerais superam BM25, e Dense 4B FT melhora o desempenho em relação aos modelos densos gerais em *NDCG@10* e *MRR@10*. Entre os reranqueadores, o melhor resultado agregado é obtido por Rerank 4B FT. Ainda assim, as melhores fusões RRF apresentam *scores* agregados superiores aos melhores sistemas densos e de reranqueamento nas métricas principais.

4.2.2. Significância estatística

Aplicamos testes pareados por consulta usando *NDCG@10* como métrica principal. Para cada comparação, calculamos o delta por consulta entre o sistema tratado e a baseline e reportamos a média desses deltas como efeito observado. Intervalos de confiança de 95% foram estimados por bootstrap pareado, enquanto a significância foi avaliada por teste de randomização aproximada unilateral ($tratamento > baseline$), com correção Holm-Bonferroni [Holm 1979] para múltiplas comparações.

Priorizamos o agregado macro estratificado por ser mais alinhado ao objetivo de comparar regimes jurídicos heterogêneos. Nesse agregado, os efeitos são calculados separadamente em cada dataset e combinados por média simples, evitando que coleções maiores dominem a conclusão.

Tabela 4. Testes pareados em *NDCG@10* no agregado macro estratificado.

Comparação	Δ médio	p Holm	Leitura
RRF 4B vs BM25	+0.0457	0.0013	significativo
RRF 8B vs BM25	+0.0553	0.0013	significativo
RRF 4B FT vs BM25	+0.0607	0.0013	significativo
RRF 0.6B FT vs BM25	+0.0189	0.0366	significativo
RRF 4B vs Dense 4B	+0.0149	0.0648	não significativo
RRF 8B vs Dense 8B	+0.0182	0.0184	significativo
RRF 4B FT vs Dense 4B FT	+0.0183	0.0308	significativo
RRF 0.6B FT vs Dense 0.6B FT	+0.0678	0.0013	significativo
RRF 4B vs Rerank 4B	+0.0114	0.1314	não significativo
RRF 8B vs Rerank 8B	+0.0144	0.0620	não significativo
RRF 4B FT vs Rerank 4B FT	+0.0121	0.1314	não significativo
RRF 0.6B FT vs Rerank 0.6B FT	+0.0571	0.0013	significativo
Melhor RRF vs melhor não-RRF	+0.0121	0.1314	não significativo

Os testes estatísticos mostram que todas as variantes RRF superam BM25 de forma estatisticamente significativa. Contra os recuperadores densos correspondentes, RRF apresenta ganho significativo em três das quatro comparações. Já a comparação com reranqueadores é mais heterogênea: apenas RRF 0.6B FT supera significativamente o reranqueador correspondente após correção. Além disso, a comparação entre o melhor RRF e o melhor sistema não-RRF não atinge significância estatística. Portanto, os resultados sustentam ganhos robustos de RRF sobre BM25 e sobre parte dos densos correspondentes, mas não uma dominância estatística universal sobre todos os reranqueadores.

4.2.3. Resultados por dataset

Os melhores sistemas variam entre os regimes avaliados. Dense 8B lidera BR-TaxQA-R; Rerank 0.6B FT lidera JUÁ-Juris; RRF domina JurisTCU e Ulysses-RFCorpus; e Nor-

Tabela 5. Melhores sistemas por dataset e por métrica (calculadas em @10.)

Dataset	Melhor NDCG	Melhor MRR	Melhor MAP	Melhor Recall
BR-TaxQA-R	Dense 8B (0.7879)	Dense 8B (0.8137)	Dense 8B (0.7148)	Dense 8B (0.8612)
JUÁ-Juris	Rerank 0.6B FT (0.2982)	Rerank 0.6B FT (0.2353)	Rerank 0.6B FT (0.2354)	Rerank 0.6B FT (0.5035)
JurisTCU	RRF 4B FT (0.4345)	RRF 4B FT (0.7462)	RRF 4B FT (0.2061)	RRF 4B FT (0.2903)
NormasTCU	RRF 4B FT (0.4092)	Rerank 4B FT (0.5414)	Rerank 4B (0.2916)	Rerank 4B (0.4370)
Ulysses-RFCorpus	RRF 8B (0.6372)	RRF 4B (0.7724)	RRF 4B/8B (0.5127)	RRF 8B (0.6359)

masTCU apresenta resultados divididos entre RRF e reranking. Esses resultados sugerem que a robustez agregada de RRF decorre de desempenho consistente em múltiplos cenários, e não de dominância uniforme em todos os subconjuntos.

5. Análise

Os resultados evidenciam que a recuperação jurídica brasileira não constitui um problema homogêneo, mas um conjunto de regimes de recuperação com demandas distintas de informação. Os achados sugerem que a eficácia relativa entre recuperação lexical, densa e híbrida depende menos do domínio jurídico em si e mais do tipo de evidência predominante em cada coleção. Cenários centrados em correspondência terminológica, citações normativas e jargões institucionais tendem a favorecer mecanismos lexicais como BM25, enquanto consultas formuladas em linguagem natural ou com baixa sobreposição lexical se beneficiam mais da generalização semântica dos modelos densos.

Nesse contexto, o melhor desempenho agregado das estratégias baseadas em *Reciprocal Rank Fusion* (RRF) reflete sua capacidade de combinar sinais complementares sem exigir calibração complexa de *scores*. Ao integrar a precisão léxico-sintática do BM25 com a capacidade de abstração semântica dos recuperadores densos, RRF produz uma configuração particularmente robusta frente à heterogeneidade estrutural do benchmark. A principal exceção ocorre em cenários de perguntas puras em linguagem natural, como no BR-TaxQA-R, em que modelos densos de maior escala (Dense 8B) apresentam compreensão semântica suficiente para reduzir o ganho marginal da evidência lexical. O corpus reduzido, com 715 documentos, também pode contribuir para esse padrão, pois em coleções pequenas o BM25 tende a produzir menos ruído e o ganho marginal da fusão pode diminuir.

Exemplos qualitativos de divergência. Inspecionamos consultas em que BM25 e o recuperador denso atribuíram posições muito distintas ao mesmo documento relevante. Em JUÁ-Juris, na consulta *"A decisão do administrador em não parcelar uma contratação deve ser obrigatoriamente precedida de estudos técnicos que a justifiquem"*: o documento `JURISPRUDENCIA-SELECIONADA-22198` aparece apenas na posição 61 do BM25, mas na posição 8 do modelo denso Qwen3-Embedding-4B FT, sendo promovido para a posição 3 pelo RRF. O trecho relevante discute "contratação por preço global", "ganho de escala" e "contratações parceladas", o que favorece a correspondência semântica em vez da repetição literal da consulta.

O padrão inverso ocorre em Ulysses-RFCorpus, na consulta *"projeto de lei que redefina o ato cooperativo"*: o documento relevante `PLP 271/2005` é recuperado na posição 4 pelo BM25, mas apenas na posição 349 pelo modelo denso Qwen3-Embedding-0.6B FT; o RRF o promove para a primeira posição. Nesse caso, a expressão jurídica

específica “ato cooperativo” favorece a busca lexical, enquanto o modelo denso se desloca para documentos semanticamente próximos, mas inadequados. Esses exemplos ilustram que o RRF preserva evidências complementares: generalização semântica quando há baixa sobreposição lexical e correspondência terminológica quando ela é decisiva.

Geração de candidatos: RRF vs. Reranqueamento. A robustez do RRF em relação aos *pipelines* de reranqueamento avaliados reside na fase de recuperação inicial. O reranqueamento clássico é fundamentalmente limitado pelo *recall* de sua primeira etapa: se o modelo esparso falha em recuperar um documento relevante devido ao descasamento de vocabulário, o reranqueador neural nunca terá a chance de avaliá-lo. O RRF contorna esse gargalo ao compor o *ranking* final a partir de duas distribuições distintas de candidatos. Isso explica por que o reranqueamento permanece forte em subconjuntos onde o BM25 já garante um bom *recall* inicial, mas perde estabilidade no cenário agregado do JUÁ frente à fusão híbrida.

O papel e os limites da adaptação ao domínio. O ajuste fino em *corpora* jurídicos fortalece visivelmente a representação semântica, o que é comprovado pelo RRF 4B FT alcançando os melhores resultados agregados em NDCG@10 e MRR@10. Contudo, os dados provam que a adaptação não elimina a necessidade da fusão léxico-densa. A especialização do modelo reduz o abismo entre a linguagem da consulta e o texto normativo, mas o BM25 continua fornecendo uma evidência ortogonal e estatisticamente significativa. Em suma, o principal mérito do RRF documentado neste trabalho não é uma dominância absoluta em todo e qualquer cenário isolado, mas a sua excepcional estabilidade estrutural frente à severa heterogeneidade da busca jurídica.

6. Conclusão e Trabalhos Futuros

Este trabalho investigou a complementaridade entre recuperação lexical e recuperação densa em recuperação jurídica brasileira heterogênea. Os resultados mostram que variantes baseadas em RRF alcançam as melhores configurações agregadas do benchmark JUÁ, produzindo ganhos robustos sobre BM25 e parte dos recuperadores densos correspondentes. Em particular, RRF 4B FT lidera em *NDCG@10* e *MRR@10*, enquanto RRF 8B obtém os maiores valores de *MAP@10* e *Recall@10*.

Ao mesmo tempo, a análise por dataset evidencia que recuperação jurídica não constitui um problema homogêneo. Consultas em linguagem natural tendem a favorecer recuperadores densos, enquanto tarefas dependentes de correspondência terminológica e citações normativas preservam a importância do sinal lexical. Nesse contexto, o principal mérito do RRF não é uma dominância universal em todos os subconjuntos, mas sua robustez frente a diferentes regimes de recuperação.

Esses resultados sugerem que a eficácia relativa entre recuperação lexical, densa e híbrida depende menos do domínio jurídico e mais da natureza do regime de recuperação predominante em cada coleção. Como trabalhos futuros, pretendemos avaliar reranqueadores mais expressivos, estratégias de fusão ponderada e análises qualitativas capazes de explicar quando evidências lexicais ou semânticas se tornam decisivas para a recuperação.

7. Limitações

Este trabalho apresenta uma avaliação empírica de métodos de recuperação em coleções jurídicas brasileiras associadas ao benchmark JUÁ. Embora os subconjuntos avaliados

cubram jurisprudência, atos normativos, documentos legislativos e perguntas tributárias, eles não representam toda a diversidade de tarefas, instituições, ramos do direito e estilos de consulta presentes na busca jurídica brasileira. Assim, os resultados devem ser interpretados dentro desse escopo experimental.

A comparação com reranqueamento é limitada ao tipo de reranqueador avaliado. Os reranqueadores utilizados baseiam-se em arquiteturas bi-encoder e atuam sobre candidatos recuperados por BM25 — abordagem computacionalmente mais eficiente, mas menos expressiva que métodos cross-encoder, como MonoT5, que podem apresentar comportamento distinto, especialmente em cenários que demandam maior modelagem de interação entre consulta e documento.

Todos os recuperadores densos avaliados pertencem à família Qwen3-Embedding. Modelos multilíngues, como mE5, e arquiteturas orientadas ao português baseadas em BERTimbau não foram avaliados. Portanto, os resultados devem ser interpretados com cautela ao extrapolar para outras arquiteturas de recuperação densa.

A análise estatística também impõe cuidados de interpretação. Os testes principais foram configurados como direcionais, avaliando a hipótese de que o tratamento supera a baseline. Além disso, diferentes formas de agregação respondem a perguntas distintas: o agregado micro atribui maior peso a datasets com mais consultas, enquanto o agregado macro estratificado dá peso comparável aos subconjuntos avaliados.

8. Declaração Ética

Este trabalho avalia métodos de recuperação de informação em documentos jurídicos, mas não propõe substituir a análise de profissionais do Direito. Sistemas de busca jurídica podem recuperar documentos incompletos, enviesados ou fora de contexto, e seus resultados devem ser interpretados como apoio à pesquisa, não como orientação jurídica definitiva. Também é importante considerar que os modelos avaliados podem reproduzir vieses presentes nos dados de treinamento ou nas coleções avaliadas. Assim, qualquer uso prático desses métodos em ambientes jurídicos deve incluir validação humana, transparência sobre suas limitações e cuidado na interpretação dos rankings produzidos.

9. Declaração de Uso de IA Generativa

Os autores utilizaram ferramentas de IA generativa de forma limitada, como apoio editorial e técnico, incluindo revisão gramatical, melhoria de clareza textual, padronização de redação e auxílio inicial na geração de trechos de código. Todo código utilizado nos experimentos foi revisado, adaptado e validado pelos autores. A concepção do estudo, a implementação final dos experimentos, a análise dos resultados, a interpretação dos achados e as conclusões são de responsabilidade integral dos autores. Todas as sugestões geradas por IA foram revisadas criticamente pelos autores antes da submissão.

10. Agradecimentos

Este trabalho foi financiado por CNPq, CAPES, FAPEMIG, FAPESP, AWS, NVIDIA e pelo Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação da Informação (INCT-TILDIAR) (processo nº 408490/2024-1).

Referências

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Chen, T., Zhang, M., Lu, J., Bendersky, M., and Najork, M. (2022). Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In *ECIR 2022*, page 95–110, Berlin, Heidelberg. Springer-Verlag.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Kamalloo, E., Thakur, N., Lassance, C., Ma, X., Yang, J.-H., and Lin, J. (2024). Resources for brewing beir: Reproducible reference models and statistical analyses. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1431–1440, New York, NY, USA. Association for Computing Machinery.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (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.

Louis, A., van Dijck, G., and Spanakis, G. (2025). Know when to fuse: Investigating non-English hybrid retrieval in the legal domain. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4293–4312, Abu Dhabi, UAE. Association for Computational Linguistics.

Nigam, S. K., Goel, N., and Bhattacharya, A. (2022). nigam@coliee-22: Legal case retrieval and entailment using cascading oflexical and semantic-based models. In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session*, page 96–108.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert.

Pereira, J., Fernandes, L., de Brito, E., Lotufo, R., and Bonifacio, L. (2026). Juá – a benchmark for information retrieval in brazilian legal text collections.

Reimers, N. and Gurevych, I. (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.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*,

pages 403–417. Springer.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (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*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models.