Concept-Based Selective Passage Indexing for Off-the-Shelf Neural Ranking in Brazilian Portuguese Legal Retrieval

Matheus dos Santos Palheta[1], **Edleno Silva de Moura**[1,2]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. Gen. Rodrigo Octávio 6200, Coroado I – 69080-900 – Manaus – AM – Brasil

{matheus.palheta, edleno}@icomp.ufam.edu.br

[2]Jusbrasil – Salvador – BA – Brasil

edleno@jusbrasil.com.br

Abstract. *This paper investigates an alternative to domain-specific fine-tuning for neural ranking models in information retrieval. We propose Concept-Based Selective Passage Indexing, a pre-encoding intervention in which each document is partitioned into sliding-window passages, and only those whose number of distinct human-curated concepts exceeds a calibrated threshold are retained for encoding. We evaluate this strategy on a legal corpus for which a high-quality set of human-curated concepts was available. Specifically, we experiment with four off-the-shelf neural rankers on a corpus of 4,268 Brazilian Federal Court decisions paired with 280 real user queries (242 in the primary evaluation set) and expert-graded relevance judgments. Across our experiments, Concept-Based Selective Passage Indexing improves NDCG relative to the full-indexing baseline for all evaluated rankers.*

Keywords: Legal Information Retrieval, Neural Ranking, Late Interaction, ColBERT, Zero-Shot Retrieval, Brazilian Portuguese, Low-Resource IR.

1. Introduction

Legal professionals rely on retrieval systems to locate relevant case law, statutes, and precedents to ground their argumentation, and the continuous growth in the volume of judicial documents has made effective Legal Information Retrieval (LIR) a critical enabler of legal practice [Mentzingen et al. 2023]. LIR systems must contend with characteristics that set the legal domain apart from general-purpose retrieval: specialized terminology, syntactic complexity, jurisdiction-dependent interpretation, and the substantial length and inherent verbosity typical of judicial decisions.

The retrieval requirements extend well beyond surface-form matching: a single legal concept admits multiple lexical realizations across statutes, doctrine, and judicial style, and the information need expressed in a user query is seldom co-located with the exact terms used in the documents that address it. Contextualized neural ranking models [Nogueira and Cho 2019, Nogueira et al. 2019, Guo et al. 2022] were introduced to capture these semantic alignments, encoding queries and documents into dense representations that score relevance beyond term overlap.

Extending these advances to non-English legal retrieval is challenging. Neural ranking models trained predominantly on English corpora exhibit limited zero-shot

generalization across languages and domains [Thakur et al. 2021]. Domain-adapted variants such as LEGAL-BERT [Chalkidis et al. 2020] consistently outperform general-purpose backbones on legal NLP tasks, underscoring the value of in-domain adaptation. Coupled with the data scarcity that characterizes most non-English legal corpora [Chari et al. 2025], this leaves practitioners in a constrained position: domain-specific fine-tuning requires labeled data and compute budgets that are frequently unavailable, while zero-shot use of off-the-shelf encoders yields a substantial effectiveness gap.

Recent work has explored interventions that mitigate this constraint without resorting to large-scale domain fine-tuning. Some approaches target the training paradigm via cross-lingual knowledge distillation [Nguyen et al. 2025a] or self-supervised pre-training over unlabeled legal corpora [Su et al. 2025]; others operate on the text representation itself, either by tailoring the tokenizer to language-specific structures [Atuhurra et al. 2024] or by filtering the input to expose only its salient content [Zhou et al. 2023].

This work investigates a pre-encoding intervention for off-the-shelf neural ranking models in Brazilian Portuguese LIR. We propose *Concept-Based Selective Passage Indexing*, an indexing strategy that restricts the encoded content of each document to passages containing human-curated legal concepts extracted from real user query logs. The method is applied exclusively at indexing time, leaving model parameters and tokenization pipelines untouched.

Our central hypothesis is that filtering out concept-absent passages before encoding allows neural ranking models to concentrate their relevance signals on semantically salient content – substantive legal reasoning rather than procedural enumerations, personnel lists, tabular appendices, or minimal-content dispositive sections (one-sentence rulings) – thereby partially compensating for the absence of domain-specific fine-tuning.

This work addresses the following research questions:

- **RQ1.** Can a passage-selection filter built from human-curated legal concepts enhance the effectiveness of off-the-shelf neural ranking models in Brazilian Portuguese LIR?
- **RQ2.** How do off-the-shelf multilingual neural rankers compare to Portuguese-specialized counterparts under both indexing conditions in this specialized domain?
- **RQ3.** To what extent can legal concepts extracted from real user query logs compensate for the absence of domain-specific training data in low-resource legal retrieval?

2. Related Work

Neural ranking architectures embody a trade-off between fine-grained query–document interaction and the ability to precompute document representations offline. Early BERT-based rankers [Nogueira and Cho 2019, Nogueira et al. 2019] used a *cross-encoder* formulation, jointly encoding the query and document with a shared transformer [Dai and Callan 2019].

This delivers strong effectiveness but requires query–document inference and incurs cost proportional to the candidate set size [Yates et al. 2021]. *Bi-*

encoders [Lu et al. 2020] decouple indexing from retrieval by encoding each document offline into a single dense vector. The *late interaction* paradigm of ColBERT [Khattab and Zaharia 2020] recovers token-level matching in this offline-indexable regime by retaining one embedding per token.

These architectures were designed and evaluated primarily on English-language benchmarks, and [Thakur et al. 2021] documented their degradation under cross-domain and cross-lingual zero-shot conditions. Two broad responses have emerged. The first is multilingual encoders pre-trained jointly over dozens of languages, exemplified by EmbeddingGemma [Vera et al. 2025] in the bi-encoder family and ColBERT-XM [Louis et al. 2025a] extending this logic to late interaction via *cross-lingual knowledge distillation*. The second is language-specific pre-training, illustrated for Brazilian Portuguese by BERTugues [Mazza Zago and Agnoletti dos Santos Pedotti 2024] and by ColBERT-BR [Berthélémé 2024], which couples this backbone with mMARCO-PT supervision.

The legal domain sharpens these challenges. Judicial decisions, statutes, and procedural documents combine specialized terminology, jurisdiction-dependent interpretation, and stylistic conventions that diverge from the general-domain corpora on which neural rankers are typically pre-trained. Labeled query–document data is also scarce for most jurisdictions and is largely absent in non-English settings [Chari et al. 2025]. These limitations are reflected in prior results: ColBERT encoders applied directly to Vietnamese legal retrieval [Tien et al. 2024] and to French statutory retrieval [Louis et al. 2025b] exhibit limited generalization without in-domain adaptation. On English data, [Chalkidis et al. 2020] established the broader principle that further pre-training on legal corpora yields substantial gains over general-domain encoders.

Two complementary lines of research mitigate this constraint without large-scale supervised fine-tuning. The first targets the training paradigm via self-supervised pre-training [Su et al. 2025], cross-lingual training [Nguyen et al. 2025b], and knowledge distillation [Nguyen et al. 2025a]; however, these approaches still incur training cost. The second operates on text representation while leaving encoder parameters untouched, including tokenizer adaptation to language-specific morphology [Atuhurra et al. 2024] and passage-level selection prior to encoding that concentrates the encoder's capacity on relevance-bearing content [Zhou et al. 2023].

Pre-encoding interventions of this kind have been explored predominantly on the query side, where additional inference cost is incurred per query at retrieval time. The present work shifts this effort to the document side. Rather than rewriting queries through generative inference, we curate the indexed content of each document by retaining only passages containing legal concepts extracted from real user query logs. This shift has two consequences. First, the intervention runs entirely offline at indexing time and incurs zero query-time latency, independent of query volume. Second, the salience signal becomes a closed, expert-curated inventory grounded in observed user behavior, replacing the open-ended generative inference characteristic of query-side approaches with a deterministic and reproducible passage-selection criterion, formalized in Section 3.

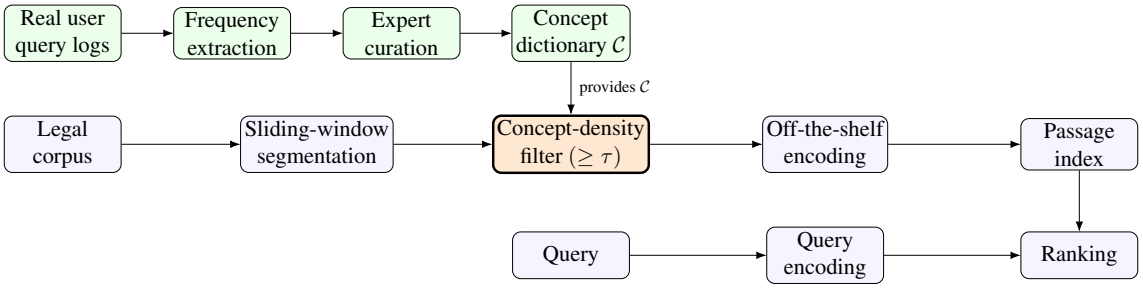


Figure 1. Overview of Concept-Based Selective Passage Indexing. A concept dictionary $\mathcal{C}$ is built offline from real user query logs (top). At indexing time, each document is segmented into sliding-window passages and only passages containing at least τ distinct concepts from $\mathcal{C}$ are encoded (middle). Retrieval (bottom) is unchanged: the query is encoded once and scored against the concept-filtered index, so the intervention incurs zero query-time overhead.

3. Methodology: Concept-Based Selective Passage Indexing

This section presents the methodology of the work. Figure 1 gives an overview of the complete system, which comprises three tracks: an *offline* track (top), in which a dictionary of legal concepts $\mathcal{C}$ is distilled from real user query logs; an *indexing* track (middle), in which each document is segmented and filtered against $\mathcal{C}$ before encoding; and a *retrieval* track (bottom), left unchanged with respect to an off-the-shelf setup.

The central design decision is already legible in the diagram: the proposed intervention is confined to a single node, the concept-density filter interposed between segmentation and encoding (shaded), while every other component is shared with the baseline. Any observed effect is therefore attributable to passage selection in isolation.

3.1. Indexing Pipeline

Brazilian Federal Court decisions frequently exceed several thousand tokens, well beyond the 512-token input limit of the BERT-based encoders evaluated in this work. Each document is therefore decomposed into a sequence of shorter passages via a sliding-window strategy. Given a document represented as a token sequence $D = (t_1, t_2, \ldots, t_N)$, the segmentation produces passages $\{P_1, \ldots, P_M\}$, each corresponding to a contiguous window of $w = 512$ tokens with a stride of $s = 462$ tokens. The resulting 50-token overlap between adjacent passages ensures that semantic units fragmented at window boundaries appear in their entirety in at least one neighboring passage. This segmentation step is applied identically under both indexing conditions defined below.

Full-Passage Indexing (FULL). The encoder processes every passage produced by the segmentation step, with no filter applied between segmentation and encoding. The resulting index contains a representation of the entire document content, distributed across overlapping passages.

Concept-Based Selective Passage Indexing (SELECTIVE). Let $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ denote a finite dictionary of legal concepts (Section 3.2). A passage P_i is retained for

encoding if and only if it contains at least τ distinct concepts drawn from $\mathcal{C}$:

$$P_i \in \text{Index} \iff \left|\{\, c_j \in \mathcal{C} : c_j \subseteq P_i \,\}\right| \geq \tau. \tag{1}$$

Here, $c_j \subseteq P_i$ denotes that the concept string c_j occurs in passage P_i. Passages failing this distinct-concept threshold are discarded prior to encoding and consequently consume no index storage nor contribute candidate matches at retrieval time.

In our experiments we set $\tau = 5$, an operating point empirically calibrated to retain substantive legal passages while discarding structurally non-substantive content. We performed experiments with a small set of 50 documents to study the value of τ.

The dictionary $\mathcal{C}$ is parameterized by query-log frequency into three operating points (Section 3.3), yielding a corresponding family of SELECTIVE indices. A central design property of the filter is that it operates at the *passage* level rather than at the *document* level: a single document typically contributes both retained and discarded passages to the index, so only rarely does a document lose all of its passages and become unindexable. The intervention reshapes *which fragments* of each document the encoder is exposed to, not *which documents* are searchable.

3.2. Concept Dictionary

The dictionary $\mathcal{C}$ was constructed from the query logs of the *Jusbrasil* platform, a Brazilian legal information service whose users are predominantly legal professionals issuing genuine information needs. A candidate set of concept strings was first extracted by computing term frequencies over the full query log: multi-word expressions and single-term strings were ranked by occurrence count, and the most frequent candidates were retained.

The resulting list spans substantive law terms (*dano moral*, *alimentos*, *prescrição*), procedural concepts (*penhora*, *citação*, *contestação*), and named legal actions (*mandado de segurança*, *ação de despejo*); Table 1 reports a sample drawn from across the frequency distribution. Candidate frequencies span more than an order of magnitude even within the high-salience region of the list, with a continuous population of intermediate candidates omitted from the table for brevity.

Table 1. Examples of legal concept candidates extracted from *Jusbrasil* query logs with their respective frequency, prior to expert curation.

Concept (PT)	Frequency	Concept (PT)	Frequency
ação	133,094	honorários	57,193
contrato	81,064	mandado de segurança	50,609
dano moral	76,768	negativação	5,882
contestação	72,493	consignação em pagamento	5,877
penhora	60,136	quebra de sigilo bancário	2,513

The full candidate list was then reviewed by legal domain experts, who discarded strings that, despite high surface frequency, lack discriminative legal meaning in isolation. The curation step removes the false positives of the frequency-ranked list and preserves only candidates that pair statistical salience with legal informativeness. The resulting dictionary $\mathcal{C}$ therefore reflects both *statistical salience* (frequency in practitioner queries) and *semantic discriminativeness* (expert judgment that the term is informative for distinguishing relevant from irrelevant documents).

3.3. Frequency-Based Dictionary Cuts

The query log itself encodes a signal beyond expert curation: the frequency with which each concept appears as a practitioner information need. To exploit this signal, we parameterize $\mathcal{C}$ by frequency bands $\mathcal{C}_{[\ell,u]}$ and define three operating points: SEL-ALL ($\mathcal{C}_{[0,\infty)}$, the entire curated dictionary), SEL-MID ($\mathcal{C}_{[5{,}000,\,50{,}000]}$, excluding both the long tail and the high-frequency core), and SEL-FREQ ($\mathcal{C}_{[10{,}000,\,\infty)}$, retaining the high-frequency core and the upper portion of the mid band).

Each cut induces its own indexed corpus: the passage-level retention criterion of Equation 1 is re-evaluated against the active $\mathcal{C}_{[\ell,u]}$, and the passages are encoded into a dedicated index. The cuts are *not strictly nested* – MID excludes the high-frequency tail admitted by FREQ, and FREQ excludes the long tail admitted by MID – so comparing across cuts constitutes an ablation over the frequency-defined regions of $\mathcal{C}$ rather than a monotonic relaxation of the filter.

Table 2. Index size and retention rate by indexing condition. The denominator is the $85{,}782$ passages produced by sliding-window segmentation under FULL.

Condition	Active dictionary	Passages indexed	Retention
FULL	—	85,782	100.00%
SEL-ALL	$\mathcal{C}_{[0,\infty)}$	84,994	99.08%
SEL-MID	$\mathcal{C}_{[5{,}000,\,50{,}000]}$	68,230	79.53%
SEL-FREQ	$\mathcal{C}_{[10{,}000,\,\infty)}$	69,102	80.55%

Retention statistics. Table 2 reports the index size and retention rate per cut against the FULL baseline of $85{,}782$ segmented passages. SEL-ALL retains 99.08% of the corpus: even under the unrestricted curated dictionary, a small minority of passages (788, or 0.92%) fails the density threshold $\tau = 5$. The two effective cuts discard substantially more – 20.47% (SEL-MID) and 19.45% (SEL-FREQ) – with the discards distributed across the corpus rather than concentrated in a pathological subset of documents: under SEL-MID, approximately $2{,}170$ of the $4{,}268$ documents lose at least one passage and only 104 lose all passages and are therefore dropped from the index.

The median concept count per passage is 12, more than double τ – the retained majority exceeds the threshold by a comfortable margin while the discarded minority concentrates well below it, indicating two structurally distinct passage populations rather than a noisy continuum.

3.4. Illustrative Filter Behavior

To make the filter concrete, consider a civil public inquiry report on the enforcement of a statutory disability-hiring quota (Law 8.213/1990). Sliding-window segmentation yields twelve passages: the first six span the procedural header, the quota computation, the substantive reasoning, and the archiving decision; the remaining six enumerate names of hired employees. Table 3 aggregates the resulting SEL-MID outcomes.

The filtering boundary aligns closely with the structural transition from substantive legal argumentation to nominal enumeration. Under FULL, a query such as

Table 3. Aggregated SEL-MID outcomes ($\tau = 5$) for the illustrative document. Substantive passages systematically exceed the threshold; personnel enumerations fall decisively below it.

Content type	Passages	Concepts (range)	Decision
Substantive legal reasoning	6	9–12	KEEP
Personnel enumerations	6	1–3	DISCARD

"contratação de pessoas com deficiência e cota legal" must compete against six index entries whose content is a list of proper names; a late-interaction ranker may assign meaningful token similarity to surface-similar name fragments in these entries. SELECTIVE eliminates these entries while preserving every substantive passage. Analogous patterns hold for three further classes of structurally non-substantive content observed in the corpus: interleaved CPF enumerations in administrative tribunal decisions, numerically dense audit appendices, and minimal-content decisions consisting of a one-sentence dispositive section (ruling) without reasoning.

4. Results

4.1. Experimental Methodology and Protocol

We use $4{,}268$ Brazilian Federal Court decisions from the *Jusbrasil* platform, paired with 280 real user queries and expert relevance judgments on a 1–3 graded scale. The assessment pool was built with a keyword-based retriever, so documents it never surfaced default to irrelevant – a threat revisited in Section 4.5.

To control for assessment density we report three *nested* configurations: **QT0** (all 280 queries); **QT1** ($n = 254$), those with at least three assessed documents; and **QT2** ($n = 242$), the subset of QT1 with at least three documents judged relevant above the lowest grade. QT2 is the primary set, since fewer than three genuinely relevant documents yield an unstable NDCG estimate; QT0 and QT1 test robustness as density increases (Section 4.5).

NDCG is computed at $k \in \{1, 5, 10, 20\}$, primary cutoff @10, with significance from the two-sided Wilcoxon signed-rank test on paired per-query distributions ($^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$).

Four off-the-shelf rankers are evaluated under each condition: **EmbeddingGemma** [Vera et al. 2025], a multilingual bi-encoder distilled from a generative backbone; **ColBERT-BR** [Berthélémé 2024], a Portuguese late-interaction ranker fine-tuned on mMARCO-PT [Bonifacio et al. 2021]; **ColBERT-XM** [Louis et al. 2025a], a multilingual late-interaction ranker; and **Legal-BERTimbau** [Souza et al. 2020], a Portuguese legal-domain bi-encoder. None is fine-tuned on the experimental corpus. All four conditions from Section 3 (FULL, SEL-ALL, SEL-MID, SEL-FREQ) are evaluated identically across systems.

4.2. Aggregate Effect of Concept-Based Selective Passage Indexing

Table 4 reports NDCG@10 on QT2 for each condition and ranker, with markers from the paired Wilcoxon test of each SELECTIVE condition against FULL.

Table 4. NDCG@10 on QT2 ($n = 242$) by indexing condition. Significance markers report the paired Wilcoxon test against FULL.

Model	FULL	SEL-ALL	SEL-MID	SEL-FREQ
EmbeddingGemma	0.6163	0.6165^{ns}	0.6283^{***}	0.6283^{**}
ColBERT-BR	0.3289	0.3305^{ns}	0.3381^{***}	0.3395^{***}
ColBERT-XM	0.2710	0.2726^{ns}	0.2813^{***}	0.2839^{**}
Legal-BERTimbau	0.3196	0.3208^{ns}	0.3289^{**}	0.3327^{***}

SEL-ALL yields no significant difference from FULL for any model: the unrestricted dictionary discards only 0.92% of segmented passages (Table 2), too small a reduction to perturb the ranking. Both effective cuts, by contrast, produce significant gains over FULL for every ranker at @10 ($p \leq 0.0024$ across the eight (model, cut) cells), with mean absolute gains of $+0.0102$ (SEL-MID) and $+0.0121$ (SEL-FREQ).

No model degrades under either effective cut at any k on QT2, so the aggregate signal is consistent in direction and significant for all four systems, supporting an affirmative answer to RQ1. The magnitude is *modest* ($|\Delta| \in [0.01, 0.03)$) for three models; Legal-BERTimbau is the outlier, with NDCG@1 deltas of $+0.0241$ ($p = 0.0245$, SEL-MID) and $+0.0331$ ($p = 0.0230$, SEL-FREQ) – the largest single delta in the experiment.

4.3. Sensitivity to the Concept Frequency Cut

The two effective cuts admit different dictionary regions: SEL-MID excludes both the long tail and the high-frequency core; SEL-FREQ retains the core and the upper mid band (Section 3.3). At NDCG@10 on QT2, SEL-FREQ ties or marginally beats SEL-MID for every model, with cross-cut deltas $\Delta_{\text{cross}} = \text{SEL-FREQ} - \text{SEL-MID}$ of $+0.0000$ (EmbeddingGemma), $+0.0014$ (ColBERT-BR), $+0.0026$ (ColBERT-XM), and $+0.0038$ (Legal-BERTimbau) – all below the *modest* threshold $|\Delta| \geq 0.01$. Two implications follow: the frequency band is a second-order decision relative to filtering at all, and the consistently positive delta indicates that the high-frequency core – terminology recurrent in real user queries – concentrates more relevance signal per passage than the mid band, consistent with the view that query logs themselves identify the concepts most likely to mediate retrieval-time matching.

The preference is not uniform along the ranking: ColBERT-XM reverses at NDCG@20 (SEL-MID 0.3291 vs. SEL-FREQ 0.3247, $\Delta_{\text{cross}} = -0.0044$) while preferring SEL-FREQ at @1, @5, and @10. Under late interaction, mid-band terminology appears to surface additional relevant documents deeper in the ranking that the narrower SEL-FREQ dictionary discards; the other three models are flat or weakly positive at @20, so this stands as an isolated architecture–cut interaction.

4.4. Cross-Model and Cutoff Patterns

The shape of the gain along the ranking varies across systems. Figure 2 reports the per-model profile of $\Delta\text{NDCG@}k = \text{SELECTIVE} - \text{FULL}$ under both effective cuts on QT2.

Two models show non-flat profiles (Figure 2). Legal-BERTimbau is front-loaded, its SEL-FREQ gain peaking at $+0.0331$ at @1: the filter reshapes the top of its ranking rather than lifting it uniformly. ColBERT-XM diverges between cuts: SEL-MID grows

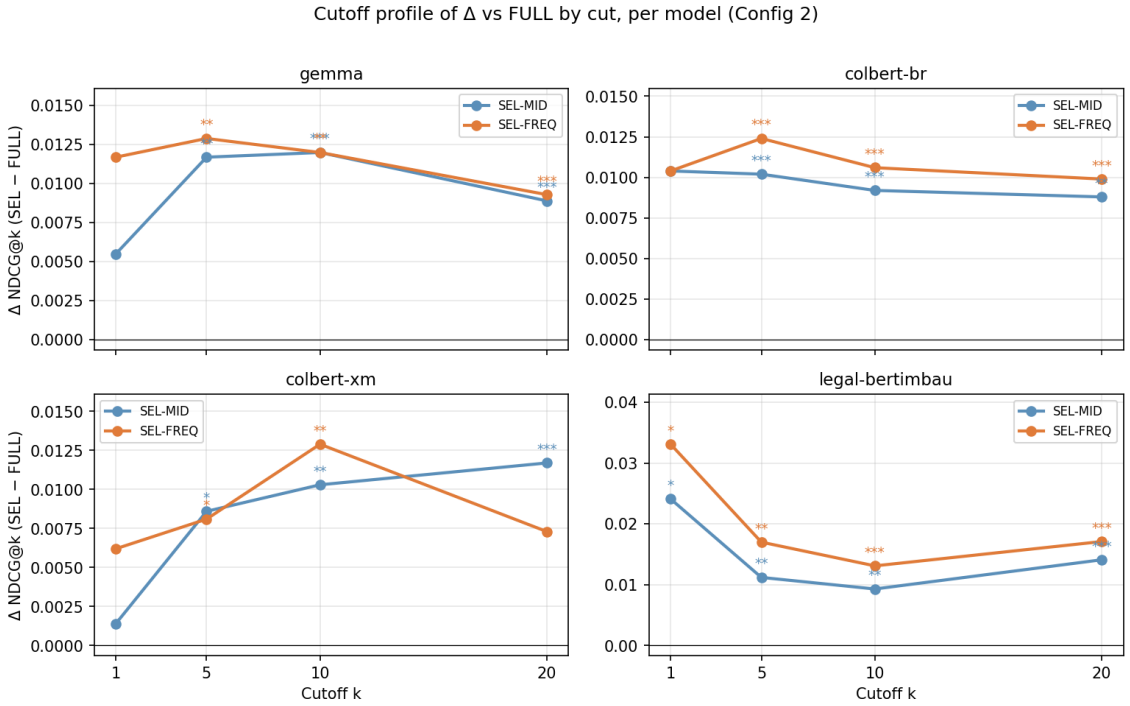


Figure 2. Cutoff profile of the SELECTIVE gain on QT2 ($n = 242$) for the two models with non-flat profiles. Legal-BERTimbau (left) is front-loaded, peaking at NDCG@1; ColBERT-XM (right) diverges between cuts, SEL-MID growing monotonically with k while SEL-FREQ peaks at @10. EmbeddingGemma and ColBERT-BR are flat under both cuts and are omitted. Asterisks: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

monotonically ($+0.0014$ at @1 to $+0.0117$ at @20), while SEL-FREQ peaks at @10 and falls to non-significance at @20. EmbeddingGemma and ColBERT-BR are flat under both cuts.

Two contrasts emerge across mean deltas. Bi-encoders extract more lift than late-interaction models ($+0.0121$ vs. $+0.0088$ under SEL-MID; $+0.0158$ vs. $+0.0097$ under SEL-FREQ), consistent with the mechanism: architectures that compress a document into one representation benefit more when the input is already curated for salient content. Portuguese-pre-trained models likewise gain more than multilingual ones. With $n = 2$ per group, both contrasts answer RQ2 only suggestively.

4.5. Robustness Across Query Configurations and Threats to Validity

Table 5 reports each effective cut's NDCG@10 delta against FULL across the three configurations, which progressively remove queries with sparse or weakly graded judgments.

Seven of eight (model, cut) trajectories strengthen monotonically as the set becomes more restrictive; EmbeddingGemma under SEL-MID is the lone exception, peaking at QT1. This is the opposite of an assessment artifact: had noisy judgments driven the effect, the gain would contract as low-quality queries are removed, not expand. The effect is therefore intrinsic to the filtering mechanism, not a function of assessment density.

A second threat is that the gain reduces to candidate-set size, since SEL-MID and SEL-FREQ discard 20.47% and 19.45% of passages (Table 2). Three observations make

Table 5. Robustness of the SELECTIVE gain at NDCG@10 across query configurations.

Model	Cut	Δ QT0	Δ QT1	Δ QT2
EmbeddingGemma	SEL-MID	+0.0115	+0.0127	+0.0119
EmbeddingGemma	SEL-FREQ	+0.0103	+0.0113	+0.0119
ColBERT-BR	SEL-MID	+0.0080	+0.0088	+0.0092
ColBERT-BR	SEL-FREQ	+0.0092	+0.0101	+0.0106
ColBERT-XM	SEL-MID	+0.0089	+0.0099	+0.0104
ColBERT-XM	SEL-FREQ	+0.0112	+0.0123	+0.0129
Legal-BERTimbau	SEL-MID	+0.0080	+0.0088	+0.0092
Legal-BERTimbau	SEL-FREQ	+0.0077	+0.0124	+0.0131

this unlikely: the two cuts shrink the set by nearly identical amounts (1.02 pp) yet yield distinguishable profiles (the ColBERT-XM @20 reversal); per-model deltas under a fixed cut vary substantially, incompatible with a size-driven uniform lift; and SEL-ALL produces no significant gain, confirming the metadata pipeline alone does not move NDCG. The gain thus reflects the concept-driven content of the retained passages rather than candidate-set reduction.

Finally, the keyword-based pool (Section 4.1) excludes any document a lexical retriever would not surface, penalizing systems that rank genuinely relevant but lexically dissimilar documents. We therefore read the absolute NDCG figures as lower bounds on the true effectiveness of the neural systems.

5. Conclusion

We introduced *Concept-Based Selective Passage Indexing*, a pre-encoding intervention that indexes only passages whose count of distinct human-curated concepts, mined from real user query logs, exceeds a calibrated threshold. Leaving parameters, tokenizers, and retrieval-time scoring untouched, it deploys on any off-the-shelf neural ranker and, on 4,268 Brazilian Federal Court decisions, produced statistically significant NDCG@10 gains over FULL for all four rankers under both effective cuts, with no degradation across the panel's 32 cells (**RQ1**).

Bi-encoders extracted more lift than late-interaction architectures, and Portuguese-pre-trained models more than multilingual ones, though $n = 2$ per group renders these contrasts suggestive rather than conclusive (**RQ2**). The robustness analysis attributes the gain to the concept-driven content of the retained passages rather than to candidate-set reduction, and the effect strengthens under denser ground-truth signal, indicating that log-grounded concepts supply a measurable portion of the salience signal that domain-specific fine-tuning would otherwise provide (**RQ3**).

The cost profile is favorable: overhead is incurred once, at indexing time, and amortized across all subsequent queries – a property that distinguishes the document-side route from query-side reformulation. The strategy generalizes to any verbose-document domain for which a curated concept inventory can be derived from query logs or domain ontologies, with medical, financial, and patent retrieval as the most immediate candidates; the dictionary itself, here a fixed expert artifact, could further adapt to drift in practitioner

information needs as the underlying log grows.

6. Acknowledgments

This research is partially supported by CNPq under INCT TILD-IAR (408490/2024-1) and individual grant Edleno Moura (310573/2023-8); by the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES) financial code 001; and by FAPEAM under the POSGRAD 2026 Program.

References

Atuhurra, J., Shindo, H., Kamigaito, H., and Watanabe, T. (2024). Introducing syllable tokenization for low-resource languages: A case study with swahili. *arXiv preprint arXiv:2406.15358*.

Berthélémé, A. (2024). Colbertv1.0-bertugues-base-portuguese-mmarcopt. `https://huggingface.co/AdrienB134/ColBERTv1.0-BERTugues-base-portuguese-mmarcoPT`.

Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., and Nogueira, R. (2021). mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

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*.

Chari, A., MacAvaney, S., and Ounis, I. (2025). Improving low-resource retrieval effectiveness using zero-shot linguistic similarity transfer. In *European Conference on Information Retrieval*.

Dai, Z. and Callan, J. (2019). Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 985–988.

Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R., and Cheng, X. (2022). Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems*, 40.

Khattab, O. and Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Louis, A., Saxena, V. K., van Dijck, G., and Spanakis, G. (2025a). Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Louis, A., van Dijck, G., and Spanakis, G. (2025b). Know when to fuse: Investigating non-english hybrid retrieval in the legal domain. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Lu, W., Jiao, J., and Zhang, R. (2020). Twinbert: Distilling knowledge to twin-structured compressed bert models for large-scale retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Mazza Zago, R. and Agnoletti dos Santos Pedotti, L. (2024). Bertugues: A novel bert transformer model pre-trained for brazilian portuguese. *Semina: Ciências Exatas e Tecnológicas*.

Mentzingen, H., António, N., and Bacao, F. (2023). Automation of legal precedents retrieval: Findings from a literature review. *International Journal of Intelligent Systems*.

Nguyen, C. M., Tran, H. A. P., Thai, P. T., Nguyen, H. D., and Nguyen, V. T. (2025a). Lightweight semantic search for low-resource languages: A case study in vietnamese information retrieval. In *International Conference on Similarity Search and Applications*. Springer.

Nguyen, T. N., Hai, N. L., Hieu, N. D., Nguyen, D. A., Van, L. N., Nguyen, T. H., and Dinh, S. (2025b). Improving Vietnamese-English cross-lingual retrieval for legal and general domains. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Association for Computational Linguistics.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Nogueira, R. F., Yang, W., Cho, K., and Lin, J. (2019). Multi-stage document ranking with BERT. *CoRR*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Su, W., Ai, Q., Wu, Y., Xie, A., Wang, C., Ma, Y., Li, H., Wu, Z., Liu, Y., and Zhang, M. (2025). Pre-training for legal case retrieval based on inter-case distinctions. *ACM Transactions on Information Systems*.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Tien, S. P., Doan, H. N., Dai, A. N., and Viet, S. D. (2024). Improving vietnamese legal document retrieval using synthetic data. *arXiv preprint arXiv:2412.00657*.

Vera, H. S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S. R., Smoot, S., Naim, I., Zou, J., Chen, F., et al. (2025). Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Yates, A., Nogueira, R., and Lin, J. (2021). Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

Zhou, Y., Huang, H., and Wu, Z. (2023). Boosting legal case retrieval by query content selection with large language models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23. Association for Computing Machinery.