Small Language Models for Legal RAG in Portuguese: A 4×2 Cross-Domain Evaluation of Faithfulness, Retrieval Strategies, and Model Scale

Guilherme Dutra[1], **André Caraíba**[1], **Nádia Felix**[1], **Daniel Ribeiro**[1], **Állan Silva**[1], **Paulo Victor dos Santos**[1], **Pedro Albernaz**[1], **Sávio Teles**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

guilhermecorreiadutra@gmail.com, nadia@inf.ufg.br

Abstract. *We present a 4×2 factorial evaluation of Retrieval-Augmented Generation (RAG) for Brazilian Portuguese legal text across four sub-domains of the JUÁ benchmark—Ulysses-RFCorpus (legislative), JurisTCU (judicial), BR-TaxQA (tax), and NormasTCU (regulatory). We compare BM25 Okapi and BGE-M3 retrieval across eight Small Language Models (SLMs, 1B–31B), with faithfulness scored by a Llama-3.3-70B judge against a Qwen-2.5-72B Oracle and validated via Cohen's κ. Whereas prior work established that retrieval effectiveness is domain-dependent at the retrieval level, we provide the first end-to-end characterization of how this is reflected in generation faithfulness: BM25 dominates legislative, judicial, and regulatory text while dense retrieval wins only tax law, and 7–8B SLMs match models up to 31B with no detectable faithfulness difference. We further identify a truncation boundary where whole-document dense retrieval collapses on ultra-long norms, and a* BERTScore Paradox *that cautions Oracle-referenced evaluation. These results argue that domain-aware retrieval paired with 7–8B SLMs—rather than larger models—is the most defensible deployment path for Brazilian legal RAG. We report associations under fixed retrieval conditions and discuss the limits of a causal reading in Section 6.*

1. Introduction

Brazilian legal information systems operate at substantial scale: the Chamber of Deputies maintains over 105,000 legislative proposals, the jurisprudence portal of the Tribunal de Contas da União (TCU, the Brazilian Federal Court of Accounts) receives 1.5M+ queries per year [Fernandes et al. 2025], and the tax code spans hundreds of interlinked normative documents—yet automated answer generation across the full breadth of Brazilian legal sub-domains remains poorly characterized.

Retrieval-Augmented Generation (RAG) [Lewis et al. 2020] has become the dominant paradigm for knowledge-intensive NLP, decoupling factual knowledge from model parameters—a decisive advantage for legal systems whose corpora evolve daily.

Since 2024, a new wave of Small Language Models (SLMs) [Belcak et al. 2025] has shifted the cost-quality frontier. Llama 3.2 [Dubey et al. 2024], Gemma 3 and 4 [Gemma Team, Google 2025, Gemma Team, Google 2026], Qwen3 [Qwen Team 2025], and Phi-4 [Abdin et al. 2024] reach competitive performance at a fraction of larger-model cost. *Yet the performance of these recent*

architectures across the distinct textual topologies of Brazilian legal sub-domains remains uncharacterized.

The JUÁ benchmark [Pereira et al. 2026] recently introduced a unified evaluation framework for Brazilian legal retrieval across six sub-domains. It established that retrieval effectiveness is domain-dependent—BM25 stays highly competitive across most collections, with dense retrieval gaining ground only in specific settings—but evaluates retrieval only, leaving unexplored whether this domain dependence carries over to generation faithfulness in a RAG setting. Our work directly addresses this gap: we adopt four JUÁ sub-domains and extend evaluation to end-to-end RAG with eight SLMs, measuring how retrieval differences relate to downstream answer quality.

We address four research questions; each labeled finding in Section 5 is tagged with the question it answers. **RQ1:** How does retrieval paradigm (BM25 vs. dense zero-shot) interact with domain to shape generation faithfulness? **RQ2:** Can SLMs ($\leq$10B) match larger models in faithfulness at lower cost? **RQ3:** How reliable is LLM-as-judge faithfulness across these sub-domains, and how should its scores be interpreted? **RQ4:** What are the failure modes and safety boundaries of zero-shot dense retrieval on legal corpora?

This paper makes three contributions:

1. **4$\times$2 cross-domain RAG benchmark (RQ1).** First end-to-end characterization of how the retrieval-paradigm inversion is reflected in generation faithfulness across four Brazilian legal sub-domains, extending JUÁ's retrieval-only findings [Pereira et al. 2026] (associations under fixed retrieval conditions; causal limits discussed in Section 6).
2. **8-model SLM benchmark (RQ2).** First evaluation of eight SLMs (1B–31B) including frontier 8B-class models against a 72B Oracle for Portuguese legal RAG.
3. **Failure-boundary documentation (RQ4).** A truncation-induced retrieval failure of zero-shot BGE-M3 on ultra-long regulatory documents, establishing a practical safety boundary and motivating chunk-level retrieval (Section 6).

2. Related Work

Our study sits at the intersection of four research threads: retrieval-augmented generation, small language models, the lexical-versus-dense retrieval debate, and legal information retrieval in Portuguese. We review each in turn and, for each, make explicit the specific prior result that our experiments build on or quantify.

Retrieval-Augmented Generation. RAG [Lewis et al. 2020] couples generation to retrieved evidence. Retrieval quality and faithfulness are sublinearly correlated [Salemi and Zamani 2024], and SLMs below 7B can fail to utilize retrieved evidence even under oracle-quality context [Pandey et al. 2026]—both patterns we quantify.

Small Language Models. SLMs ($\leq$10B) are architecturally suited for modular RAG pipelines [Belcak et al. 2025], reaching competitive quality at a fraction of larger-model cost while remaining deployable on modest hardware.

Retrieval: lexical, dense, and hybrid. BM25 Okapi [Robertson and Zaragoza 2009] is the canonical sparse baseline. BGE-M3 [Chen et al. 2024] achieves strong zero-shot cross-lingual dense retrieval (570M params, 8192-token limit); hybrid fusion via reciprocal rank fusion (RRF) typically outperforms both components. We additionally include multilingual MiniLM-L12-v2 (118M) as a compact dense baseline aligned with prior Brazilian legal IR [Fernandes et al. 2025]. Long-chunk truncation produces hard-negative contamination [Jiang et al. 2024]—directly relevant to our NormasTCU failure.

Legal IR in Portuguese. Brazilian legal IR has matured dataset by dataset, but each prior resource targets a single sub-domain and stops at retrieval. Vitório *et al.* [Vitório et al. 2025] released Ulysses-RFCorpus and Souza *et al.* [Souza et al. 2021] established BM25L as its baseline; Fernandes *et al.* [Fernandes et al. 2025] introduced JurisTCU, finding that BERT-scale dense models underperform BM25 on TCU queries. JUÁ [Pereira et al. 2026] unified six sub-domains into the first multi-domain Brazilian legal retrieval benchmark, confirming that retrieval effectiveness is domain-dependent but without RAG evaluation. We build directly on this line, extending the question from retrieval through to end-to-end generation.

Automatic RAG evaluation. RAGAS defines Faithfulness as the fraction of claims supported by context [Es et al. 2024], the primary metric we adopt. LLM-as-judge protocols are validated [Kim et al. 2024, Shi et al. 2025], with circular-bias risks when Oracle and Judge share a model family [Liu et al. 2025]. Frontier-LLM-generated references are established as valid anchors [Zhang et al. 2024].

3. Datasets

We evaluate on four JUÁ sub-domains chosen to span a wide range of textual topologies—from short, abstract-like legislative summaries to ultra-long regulatory norms—so that retrieval and generation can be compared across genuinely different document structures. **Ulysses-RFCorpus** [Vitório et al. 2025] contains 105,681 legislative bills and 668 advisor-authored queries (graded relevance by 54 specialists); we index the `txt_ementa` abstract ($\sim$100 words), where informal queries meet formal bureaucratic text—a persistent vocabulary gap. **JurisTCU** [Fernandes et al. 2025] has 16,045 TCU rulings and 150 queries; we index only the EXCERTO field ($\sim$660 words) for direct comparison with JUÁ [Pereira et al. 2026]. **BR-TaxQA-R** [Domingos Júnior et al. 2025] has 715 income-tax (IRPF) queries with graded passages from the Receita Federal corpus [Pereira et al. 2026] (primary answer score 2, linked FAQ score 1); its structured article vocabulary makes dense matching a credible alternative to lexical retrieval. **NormasTCU**[1] has 46 queries over normative documents averaging $\approx$12,000 tokens—far exceeding BGE-M3's 8,192-token limit and motivating the truncation analysis (Section 5.5).

4. Methodology

We now describe the retrieval systems, the eight generation models (together with the Oracle and the Judge), the RAG pipeline and its prompts, the 4×2 factorial design, and the

[1] `https://huggingface.co/datasets/ufca-llms/normas-tcu`

Table 1. The eight SLMs evaluated. The Qwen2.5-72B Oracle and the Llama-3.3-70B (primary) and Gemini-2.5-flash (κ) judges are specified in the text.

Model (abbrev.)	Params	Family	Role
Llama-3.2-1B-Instruct (Llama-1B)	1B	Meta	SLM
Llama-3.2-3B-Instruct (Llama-3B)	3B	Meta	SLM
Gemma-3-4B-IT (Gemma-3-4B)	4B	Google	SLM
Qwen2.5-7B-Instruct (Qwen2.5-7B)	7B	Alibaba	SLM
Qwen3-8B	8B	Alibaba	SLM
Ministral-8B-2512 (Ministral-8B)	8B	Mistral	SLM
Phi-4	14B	Microsoft	Mid-range
Gemma-4-31B-IT (Gemma-4-31B)	31B	Google	Upper bound

evaluation metrics. The factorial design disentangles three effects usually confounded in RAG studies: the model, the retriever, and the legal domain itself.

4.1. Retrieval Systems

We evaluate four systems. **BM25 Okapi** [Robertson and Zaragoza 2009]: rank-bm25 (k_1=1.5, b=0.75) with symmetric NFD Unicode normalization. **Dense MiniLM-384**: multilingual MiniLM-L12-v2 (118M params), the compact baseline used by Fernandes *et al.* [Fernandes et al. 2025]. **Dense BGE-M3-1024**: BAAI BGE-M3 [Chen et al. 2024] (570M params, 8192-token limit). **Hybrid BM25+BGE-M3**: RRF (k=60). All systems retrieve top-50 (dense via FAISS IndexFlatIP); the full four-system comparison appears in Table 2. **For the generation factorial we deliberately keep only BM25 and BGE-M3**, the cleanest lexical-vs-dense contrast: MiniLM is excluded because its near-zero nDCG (Table 2) would yield an uninformative generation condition, and Hybrid is excluded because, fusing both components, it cannot isolate the retriever main effect of RQ1.

4.2. Generation Models

Table 1 lists all evaluated models, with abbreviated names used in subsequent tables and text. All accessed via DeepInfra API (temperature = 0.1; 1,024 output tokens).

Oracle. Qwen-2.5-72B-Instruct generates reference responses for BERTScore and ROUGE, following the response-adapted reference paradigm of RevisEval [Zhang et al. 2024]. It is kept distinct from the Llama-family Judge to avoid circular bias [Liu et al. 2025].

Judge. Llama-3.3-70B-Instruct evaluates Faithfulness following [Es et al. 2024].

Inter-rater validation. Cohen's κ [Cohen 1960] between Llama-3.3-70B and a second Gemini-2.5-flash judge on 150 samples per experiment, following [Shi et al. 2025, Kim et al. 2024].

4.3. RAG Pipeline and Prompts

Top-K=5 documents are formatted into a structured prompt, each truncated to its first 2,500 characters — head truncation, with no query-aware passage selection — a per-document budget that caps the effective context at ≈3,000 tokens. Because document length differs sharply across sub-domains, the realized context budget is domain-dependent: the mean prompt length ranges from 467 tokens on Ulysses (short `ementa` abstracts, ≈100 words)

Table 2. Retrieval nDCG@10. †=significantly different from BM25 ($p<0.001$). ⋆=nDCG=0.000 due to architectural truncation (see Section 5.5). ‡=not evaluated on BR-TaxQA (outside 4×2 scope). Bold=best per corpus.

System	Ulysses	JurisTCU	BR-TaxQA	NormasTCU
BM25 Okapi	**0.545**	**0.375**	0.595	**0.354**
Dense MiniLM	0.190†	0.068†	—‡	0.000†⋆
Dense BGE-M3	0.311†	0.173†	**0.710**	0.000†⋆
Hybrid BM25+BGE	0.503†	0.253†	—‡	0.216†

to ≈3,860 tokens on JurisTCU (EXCERTO, ≈660 words) and NormasTCU (12k-token norms, truncated). The complete code, prompts, and evaluation pipeline are publicly available at `https://github.com/GuiiCorreia/SLM-Legal-RAG-PT`.

4.4. 4×2 Factorial Design

Each of the four corpora is evaluated under two retrieval conditions (BM25 and BGE-M3), yielding eight experiments E1a–E4b (per-condition retrieval nDCG@10 is listed in Table 2; BGE-M3 beats BM25 only on BR-TaxQA, E3b). This isolates three contrasts: *(i)* model effect (fix retriever and corpus); *(ii)* retriever effect (fix corpus, compare BM25 vs. BGE-M3); *(iii)* domain effect (fix retriever, compare corpora).

4.5. Evaluation Metrics

Retrieval: nDCG@5 and nDCG@10 [Järvelin and Kekäläinen 2002], Recall@10, MAP. Significance via Wilcoxon signed-rank (α=0.05, vs. BM25).

Generation: ROUGE-1/2/L [Lin 2004] and BERTScore F1 [Zhang et al. 2020] (multilingual BERT-base, layer 9) against Oracle; Faithfulness via LLM-as-judge (0–1) [Es et al. 2024]. Bootstrap 95% CI (1,000 iterations, seed 42); Wilcoxon signed-rank on per-query faithfulness, cross-validated against a normal-approximation implementation ($|\Delta p|<10^{-6}$); reported p-values are per-comparison, uncorrected for multiplicity, so isolated marginal results should be read as exploratory. Non-significant Wilcoxon results indicate absence of detectable difference rather than equivalence; see Limitations.

5. Results

We report retrieval first (Section 5.1: the nDCG@10 baseline that conditions all generation), then generation per sub-domain under the 4×2 design (Sections 5.2–5.5; E1: Ulysses, E2: JurisTCU, E3: BR-TaxQA, E4: NormasTCU), then cross-domain and judge-reliability analysis (Sections 5.6–5.7). Each labeled finding maps to one research question: **RQ1** (retrieval×domain) by R1, R3, G4, G5, G6, G9; **RQ2** (SLM scale) by G1, G2, G7; **RQ3** (evaluation validity) by G3 and G10; and **RQ4** (failure boundaries) by R2, G8.

5.1. Retrieval Comparison

Finding R1 – Cross-domain retrieval inversion. BM25 Okapi achieves the highest nDCG@10 on three of four corpora. BGE-M3 is the sole exception, surpassing BM25 on BR-TaxQA (0.710 vs. 0.595, +19%, $p<0.001$), extending the JUÁ [Pereira et al. 2026]

Table 3. Generation quality on Ulysses (n=668; Phi-4 n=659). Faith=Faithfulness; BF1=BERTScore F1; ΔF=BGE-M3−BM25 (positive=BGE-M3 better). †Wilcoxon $p<0.05$. Bold=best per column (excl. Oracle).

		E1a: BM25		E1b: BGE-M3		
Model	**P**	**Faith**	**BF1**	**Faith**	**BF1**	**ΔF**
Oracle	72B	1.00	1.000	1.00	1.000	—
Llama-1B	1B	0.590	0.764	0.547	0.767	−0.044†
Llama-3B	3B	0.585	0.764	0.538	0.767	−0.047
Gemma-3-4B	4B	0.708	0.769	0.668	0.776	−0.040†
Qwen2.5-7B	7B	0.802	0.706	0.763	0.714	−0.040
Qwen3-8B	8B	0.791	0.800	0.766	0.810	−0.025
Ministral-8B	8B	0.806	0.768	0.764	0.772	−0.041†
Phi-4	14B	**0.817**	**0.812**	**0.790**	**0.821**	−0.028†
Gemma-4-31B	31B	0.748	0.784	0.734	0.787	−0.015

retrieval-level findings to our experimental setup: BR-TaxQA's precise article-level vocabulary benefits dense retrieval, while Ulysses and JurisTCU queries rely on keyword matching that BM25 handles better.

Finding R2 – Truncation-induced dense-retrieval failure on NormasTCU (RQ4). Under the *naive whole-document* setting, both dense encoders collapse to nDCG@10=0.000 on NormasTCU (no relevant document in the top ranks). These documents average ≈12k tokens, far exceeding the input windows of both encoders—BGE-M3's 8192 tokens [Chen et al. 2024] and MiniLM's much shorter context—so whole-document truncation discards the passages carrying the query-relevant normative references [Jiang et al. 2024], while BM25 (nDCG=0.354) is unaffected. We evaluate the whole-document setting deliberately, to expose the truncation boundary that a naive dense deployment would hit; chunk-level indexing is the standard mitigation and we return to it in Section 6.

Finding R3 – NFD normalization explains BM25 advantage over JUÁ. Our BM25 achieves 0.545 on Ulysses (JUÁ: 0.531, +2.6%) and 0.375 on JurisTCU (JUÁ: 0.362, +3.5%) via symmetric NFD Unicode normalization removing accent mismatches.

5.2. E1 – Legislative Domain: Ulysses (668 queries)

Table 3 reports results for E1a (BM25, nDCG=0.545) and E1b (BGE-M3, nDCG=0.311).

Finding G1 – Scale collapse at ≤3B (RQ2). Llama-1B (0.590) and Llama-3B (0.585) reach near-identical Faithfulness despite the 3× parameter increase, producing short outputs (ROUGE-L≈0.35) that paraphrase rather than cite retrieved evidence [Pandey et al. 2026].

Finding G2 – A 7–14B quality plateau (RQ2). Phi-4 (0.817) leads E1a, with Ministral-8B (0.806), Qwen2.5-7B (0.802), and Qwen3-8B (0.791) within 2.6 pp; no pairwise difference among these four is significant (all Wilcoxon $p>0.06$, n=668). We do not claim formal equivalence (no TOST), but a 7–8B model delivers Faithfulness comparable to a 14B one—evidence of parameter efficiency at this scale [Belcak et al. 2025]. Gemma-4-31B (0.748) falls significantly below Phi-4 ($p<0.001$), indicating diminishing returns beyond 8B.

Table 4. Faithfulness on JurisTCU (E2, n=150) and BR-TaxQA (E3, n=715). ΔF=BGE-M3−BM25 (positive=BGE-M3 better). †Wilcoxon $p<0.05$. Bold=best per condition (excl. Oracle).

		JurisTCU (E2)			BR-TaxQA (E3)		
Model	**P**	**BM25**	**BGE-M3**	**ΔF**	**BM25**	**BGE-M3**	**ΔF**
Oracle	72B	1.00	1.00	—	1.00	1.00	—
Llama-1B	1B	0.542	0.558	+0.016	0.626	0.690	+0.064
Llama-3B	3B	0.521	0.576	+0.055	0.641	0.692	+0.051
Gemma-3-4B	4B	0.673	0.695	+0.022	0.725	0.764	+0.039
Qwen2.5-7B	7B	0.657	0.635	−0.022	0.670	0.705	+0.035
Qwen3-8B	8B	0.601	0.673	+0.073†	0.740	**0.772**	+0.032
Ministral-8B	8B	0.705	0.726	+0.021	0.740	0.762	+0.022
Phi-4	14B	0.700	0.724	+0.024	0.742	0.780	+0.038
Gemma-4-31B	31B	**0.773**	**0.755**	−0.017	0.786	**0.815**	+0.029

Finding G3 – A BERTScore Paradox (RQ3). High-Faithfulness models can rank low on Oracle-referenced BERTScore: Qwen2.5-7B is 3rd in Faithfulness (0.802) but 8th in BF1 (0.706), as its verbose 630-token outputs diverge stylistically from the concise Oracle, whereas Phi-4 leads both metrics. We analyze the mechanism in Section 6.1.

Finding G4 – BM25 uniformly dominates Ulysses generation (RQ1). BM25 beats BGE-M3 for *all* 8 models (mean ΔF=-0.035; the four per-model differences marked † in Table 3 are statistically significant), consistent with sublinear retrieval–faithfulness coupling [Salemi and Zamani 2024]: BM25's NFD-normalized tokenization captures canonical legislative terminology that zero-shot BGE-M3 does not recover.

5.3. E2 – Judicial Domain: JurisTCU (150 queries)

Table 4 (E2 columns) shows E2a (BM25, nDCG=0.375) and E2b (BGE-M3, nDCG=0.173); mean faithfulness drops ≈8pp vs. Ulysses due to lower retrieval quality.

Finding G5 – JurisTCU generation mildly favors BGE-M3 despite BM25's retrieval lead (RQ1). Although BM25 leads JurisTCU retrieval by +117% nDCG, 6 of 8 models attain higher Faithfulness with BGE-M3 (mean ΔF≈+0.02), though only Qwen3-8B is significant (Δ=+0.073, p=0.011; others $p>0.31$). The dense EXCERTO passages (∼660 words) are retrieved more reliably than their nDCG suggests [Salemi and Zamani 2024].

5.4. E3 – Tax-Law Domain: BR-TaxQA (715 queries)

Table 4 (E3 columns) shows E3a (BM25, nDCG=0.595) and E3b (BGE-M3, nDCG=0.710) [Domingos Júnior et al. 2025]—the only condition where BGE-M3 outperforms BM25 on retrieval.

Finding G6 – BGE-M3 retrieval advantage yields uniform faithfulness gains (RQ1). All models improve switching E3a→E3b (+2.2–6.4pp): on BR-TaxQA, the only subdomain where dense retrieval wins, higher retrieval quality is consistently associated with higher faithfulness [Salemi and Zamani 2024].

Finding G7 – The 8B frontier holds, but the largest model benefits most from the best retrieval (RQ2). The 7–8B plateau persists on BR-TaxQA: Qwen3-8B (0.740), Ministral-8B (0.740), and Phi-4 (0.742) are statistically indistinguishable in E3a ($p>0.05$).

Table 5. Generation quality on NormasTCU (n=46). ΔF=BGE-M3−BM25; ↓x: BGE-M3 collapses by x; ↑x: BGE-M3 gains by x. E4b: BGE-M3 nDCG=0.000 (8192-token truncation failure). Bold=best per condition (excl. Oracle).

Model	P	E4a (BM25)	E4b (BGE-M3)	ΔF
Oracle	72B	1.00	1.00	—
Llama-1B	1B	0.422	0.165	↓0.257
Llama-3B	3B	0.448	0.144	↓0.304
Gemma-3-4B	4B	0.615	0.352	↓0.263
Qwen2.5-7B	7B	0.561	0.420	↓0.141
Qwen3-8B	8B	0.528	**0.570**	↑0.042
Ministral-8B	8B	0.554	0.209	↓0.345
Phi-4	14B	0.489	0.413	↓0.076
Gemma-4-31B	31B	**0.652**	0.483	↓0.169

The one model that pulls ahead is the largest: with BGE-M3's best-in-benchmark retrieval (nDCG=0.710), Gemma-4-31B reaches 0.815 [0.803, 0.827], the highest Faithfulness across all eight experiments—larger models benefit disproportionately when, and only when, context quality is high.

5.5. E4 – Regulatory Domain: NormasTCU and Cascading Failure

Table 5 shows E4a (BM25, nDCG=0.354) and E4b (BGE-M3, nDCG=0.000).

Finding G8 – Retrieval failure propagates to a faithfulness collapse (RQ4). With BGE-M3 retrieving zero relevant documents, small models collapse: Llama-1B (16.5%), Llama-3B (14.4%), Ministral-8B (20.9%) achieve near-random Faithfulness [Pandey et al. 2026]. Ministral-8B's collapse (from 0.806 on Ulysses) shows instruction-following produces confident but hallucinated citations when retrieval fails.

5.6. Cross-Domain Faithfulness Analysis

Finding G9 – Cross-domain faithfulness tracks retrieval quality and document structure (RQ1). Averaging per-model Faithfulness under BM25 (the values in the BM25 columns of Tables 3–5), the four domains rank Ulysses (0.731) > BR-TaxQA (0.709) > JurisTCU (0.647) > NormasTCU (0.534). Two factors explain this order: retrieval quality (Ulysses' nDCG=0.545 is highest) and document structure (BR-TaxQA's article format supports extractive generation despite lower nDCG). The EXCERTO advantage on JurisTCU is thus *conditional* on sufficient retrieval quality [Salemi and Zamani 2024], and NormasTCU's small corpus (46 queries) caps its absolute ceiling.

5.7. Judge Reliability: Cohen's κ

Finding G10 – Judge agreement is domain-sensitive and reveals where Faithfulness rankings should be interpreted with care. Cohen's κ (Table 6) spans 0.04–0.65 (Slight to Substantial), validated on three models spanning the 4B–8B scale range across the four BM25 conditions. Agreement is strongest on NormasTCU (κ=0.65, Substantial, for Ministral-8B), where regulatory text admits clearer binary citation vs. hallucination judgments [Kim et al. 2024]. JurisTCU and BR-TaxQA show Fair-to-Moderate agreement (0.23–0.46). Ulysses (E1a) shows the lowest κ (0.04–0.22), but this is a base-rate artifact rather than judge disagreement: the raw observed agreement is in fact high (0.72–0.87), while both judges label the large majority of responses faithful (faithful rates 0.70–0.95).

Table 6. Cohen's κ (Llama-3.3-70B vs. Gemini-2.5-flash). Landis & Koch [Landis and Koch 1977]: $\leq$0.20 slight (s), 0.21–0.40 fair (f), 0.41–0.60 moderate (m), 0.61–0.80 substantial (S).

Model	E1a Ulysses	E2a JurisTCU	E3a BR-TaxQA	E4a NormasTCU
Gemma-3-4B	0.19 (s)	**0.46 (m)**	0.23 (f)	0.22 (f)
Qwen3-8B	0.04 (s)	0.33 (f)	0.34 (f)	0.49 (m)
Ministral-8B	0.22 (f)	**0.46 (m)**	0.32 (f)	**0.65 (S)**

Under such skewed marginals the expected chance agreement approaches the observed rate, so κ collapses toward zero—a well-known sensitivity of κ to prevalence, not evidence that the judges disagree. We therefore treat Ulysses Faithfulness as reliable in aggregate yet report it as comparative rankings, and note that larger, prevalence-balanced judge panels would yield more informative κ [Shi et al. 2025].

6. Discussion

6.1. The BERTScore Paradox in Oracle-Referenced Evaluation

The paradox of Finding G3—high-Faithfulness models ranking low on Oracle-referenced BERTScore—stems from two mechanisms [Zhang et al. 2020]: *stylistic monoculture*, since the verbose Qwen-2.5-72B Oracle is matched more closely by concise instruction-tuned models than by models that write elaborately; and *metric orthogonality*, since BERTScore measures stylistic proximity to the Oracle whereas Faithfulness measures factual grounding. The practical lesson is to report both metrics jointly and to check Oracle–model stylistic alignment before reading BERTScore as a quality signal.

6.2. The Truncation Boundary and Causal Interpretation

Two aspects deserve a careful reading: causality and the truncation boundary. Because our design fixes the retriever within each experiment, what we observe is an *association* between retrieval quality and faithfulness—cleanest where the better dense retriever lifts every model on BR-TaxQA (Finding G6) and where the NormasTCU retrieval failure drags faithfulness down across the board (G8). This is consistent with a causal mechanism, but we do not claim controlled causal evidence: the proper test is a counterfactual that holds the model fixed while injecting gold, retrieved, and random contexts, which we report as in-progress work.

The truncation boundary itself is a property of the *naive whole-document* configuration, not of dense retrieval per se. BGE-M3's 8192-token limit cannot fit NormasTCU's $\approx$12k-token norms, so truncation removes the query-relevant normative passages: the encoder still returns rankings, but the relevant documents fall outside the dense top (Recall@10=0) while lexical BM25 retrieves them. The SLM then hallucinates rather than abstains under the resulting irrelevant context [Jiang et al. 2024, Pandey et al. 2026]. This is therefore a deployment safety boundary for naive dense pipelines—mitigated by chunk-level indexing—not an intrinsic failure.

7. Conclusions

We presented a 4$\times$2 cross-domain RAG evaluation across eight SLMs and four Brazilian legal sub-domains with BM25 and BGE-M3 retrieval, extending the JUÁ retrieval

benchmark [Pereira et al. 2026] to generation. Four conclusions:

(1) **The retrieval inversion is reflected in generation, and retrieval quality is the dominant faithfulness driver (RQ1)**: BM25 wins 3 of 4 domains and BGE-M3 only tax law, and faithfulness tracks retrieval quality under fixed retrieval conditions; a controlled causal test is in progress (Section 6).

(2) **Naive whole-document BGE-M3 hits a truncation boundary on ultra-long documents (RQ4)**: zero retrieval on NormasTCU propagates to a sharp Faithfulness collapse, and chunk-level indexing is the standard mitigation.

(3) **7B–14B SLMs show no detectable Faithfulness differences on Ulysses (RQ2)**: this favors deploying a 7–8B model over 14B at far lower cost, though without equivalence testing we do not claim true equivalence.

(4) **Evaluation validity requires care (RQ3)**: the BERTScore Paradox forces both metrics to be reported jointly, and domain-sensitive judge agreement means Ulysses Faithfulness should be read as comparative rankings, not absolute scores.

Practical implication. For Brazilian public-sector legal RAG, our results indicate that a domain-aware retrieval choice paired with a 7–8B SLM is the most cost-defensible deployment configuration: comparable Faithfulness to 14–31B models at a fraction of the parameter count, with measurable safety boundaries on regulatory text.

Limitations. Our central reliability caveat is that faithfulness is scored by an LLM judge against an LLM Oracle. We mitigate this with a second judge and Cohen's κ; the low κ on legislative text reflects skewed marginals (most responses judged faithful), not judge disagreement—raw agreement stays high (Finding G10)—but we still report Ulysses faithfulness as comparative rankings rather than calibrated measurements. A human evaluation is in progress (a sample stratified across the eight E1a–E4b conditions, rated by multiple annotators with human-vs-LLM Cohen's κ), which will validate the judge and extend agreement evidence to the BGE-M3 conditions (the LLM-only κ here covers only the BM25 conditions). Beyond this, findings on NormasTCU (n=46) are exploratory; retrieval is zero-shot only, and fine-tuned embeddings [Pereira et al. 2026] are expected to narrow the dense gap; we do not run formal equivalence testing (*e.g.* TOST), so non-significant differences indicate absence of a detectable difference, not proof of equivalence; all models were run at a single low temperature (0.1) without per-model inference tuning, which may under-serve reasoning-capable models (*e.g.* Gemma-4, Qwen3) configured for higher-temperature decoding; and external validity is bounded to four legal sub-domains in Brazilian Portuguese, leaving generalization to other sub-domains (*e.g.* labor, criminal) or Romance languages open.

Future work. Future work follows directly: the human evaluation and a counterfactual context ablation (gold vs. retrieved vs. random context at a fixed model) to move from association toward causation; chunk-level dense retrieval to turn the NormasTCU truncation boundary into a working condition; and Ulysses-MTRAG, a synthetic multi-turn RAG benchmark for Brazilian legislative dialogue [Rosenthal et al. 2026], with fine-tuned legal encoders [Pereira et al. 2026] to close the retrieval gap.

Acknowledgements

This work was supported by Centro de Excelência em Inteligência Artificial. This work was also supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

The authors used large language models to assist with language and formatting; they reviewed and edited all content and take full responsibility for the final publication.

Competing interests. The authors have no relevant competing interests to declare.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., et al. (2024). Phi-4 technical report.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., and Molchanov, P. (2025). Small language models are the future of agentic AI.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Domingos Júnior, J., Faria, A., de Oliveira, E. S., de Brito, E., Teotonio, M., Assumpção, A., Carmo, D., Lotufo, R., and Pereira, J. (2025). BR-TaxQA-R: A dataset for question answering with references for Brazilian personal income tax law, including case law. In *Intelligent Systems (BRACIS 2025)*. Springer.

Dubey, A. et al. (2024). The Llama 3 herd of models.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2024)*, pages 150–158.

Fernandes, L. C., Ribeiro, L. d. S., Castro, M. V. B. d., Pacheco, L. A. d. S., and Sandes, E. F. d. O. (2025). JurisTCU: A Brazilian Portuguese information retrieval dataset with query relevance judgments. *Language Resources and Evaluation*.

Gemma Team, Google (2025). Gemma 3 technical report.

Gemma Team, Google (2026). Gemma 4 model card. `https://ai.google.dev/gemma/docs/core/model_card_4`.

Jiang, Z., Ma, X., and Chen, W. (2024). LongRAG: Enhancing retrieval-augmented generation with long-context LLMs.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., and Seo, M. (2024). Prometheus 2: An open source language model specialized in

evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4334–4353.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Liu, S., Li, X., Liu, Z., Yan, Y., Yang, C., Zeng, Z., Liu, Z., Sun, M., and Yu, G. (2025). Judge as a judge: Improving the evaluation of retrieval-augmented generation through the judge-consistency of large language models.

Pandey, S. et al. (2026). Can small language models use what they retrieve? an empirical study of retrieval utilization across model scale.

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

Qwen Team (2025). Qwen3 technical report.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Rosenthal, S., Shah, V., Katsis, Y., and Danilevsky, M. (2026). SemEval-2026 task 8 (MTRAGEval): Evaluating multi-turn RAG conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Salemi, A. and Zamani, H. (2024). Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., and Vosoughi, S. (2025). Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of IJCNLP-AACL 2025*.

Souza, E., Vitório, D., Moriyama, G., Santos, L., Martins, L., Souza, M., Fonseca, M., Félix, N., Carvalho, A., Albuquerque, H. O., and Oliveira, A. L. I. (2021). An information retrieval pipeline for legislative documents from the Brazilian Chamber of Deputies. In *Legal Knowledge and Information Systems (JURIX 2021)*, pages 119–126.

Vitório, D., Souza, E., Martins, L., da Silva, N. F. F., de Carvalho, A. C. P. d. L., Oliveira, A. L. I., and de Andrade, F. E. (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*, 59:1257–1277.

Zhang, Q., Wang, Y., Yu, T., Jiang, Y., Wu, C., Li, L., Wang, Y., Jiang, X., Shang, L., Tang, R., Lyu, F., and Ma, C. (2024). RevisEval: Improving LLM-as-a-Judge via response-adapted references.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR 2020*.