Grounded in Law: A Multi-Stage Anti-Hallucination Pipeline for Legal RAG Systems in Brazilian Portuguese

Arla Figueiredo, João Lucas, Tatiana Ribeiro, Caio Nery, Alan Rios
Caio Hebert, Luiza Florentino, Arthur Silva, Ícaro Feyerabend, Pedro Vidal and Bruno Cabral
Escavador
{arla-figueiredo, joaolucas, tatianaoliveira, caionery, alanrios}
{caiohebert, luizaflorentino, arthursilva, icarofeyerabend}
{pedrovidal, bruno}@escavador.com

Abstract

Large Language Models (LLMs) are effective text generators but create legal citations at non-trivial rates, a failure mode with serious consequences in legal practice. In Brazilian Portuguese the risk is amplified by citation variability (*juridiquês*), fragment-level references (article → paragraph → item), and the need to distinguish jurisdictions and court instances.

We describe EscavAI, a production Retrieval-Augmented Generation (RAG) system deployed at Escavador, a Brazilian legal-technology platform. The system combines (1) domain-tuned hybrid retrieval (lexical, dense, and cross-encoder reranking) over a large-scale legal corpus; (2) grounded generation with explicit citation constraints; and (3) a post-generation *Reference Audit* layer that extracts legislation and jurisprudence mentions via specialized taggers, normalizes them to a canonical schema, checks *existence* against authoritative databases at fragment granularity, verifies *fidelity* against official texts, and triggers targeted rewrites when inconsistencies are detected.

We report production telemetry from 184,895 audited answers containing 43,175 extracted legal references. Legislation references resolve at 81.7%, while jurisprudence references resolve at only 47.1%, identifying case-law normalization as the primary bottleneck for practitioners. Fidelity verification corrected 6.5% of checked answers before delivery, preventing misrepresented legal claims from reaching end users. By converting silent hallucinations into explicit warnings with per-reference status, the system enables legal professionals to trust verified citations and efficiently review flagged ones, rather than manually checking every authority.

1 Introduction

Generative AI can accelerate legal research and drafting, but factual errors in legal text carry severe consequences: citing a non-existent statute, misquoting an article, or inventing a plausible case identifier can lead to professional sanctions. In Brazil, courts have already sanctioned lawyers who submitted AI-generated texts containing fabricated jurisprudence, and recent studies report that specialized legal tools hallucinate in over 17% to 33% of responses (Magesh et al., 2025).

The problem is amplified in Brazilian Portuguese for domain-specific reasons. Legal language (*juridiquês*) contains terms with precise procedural meanings that may confuse multilingual models. Brazil's Civil Law system differs from the common-law concepts that dominate English-heavy pretraining corpora. Brazilian legal citations also require fragment-level precision (articles → paragraphs → subsections → items), making automatic verification harder than validating a single document identifier.

Retrieval-Augmented Generation (RAG) reduces hallucination by grounding output in retrieved documents (Lewis et al., 2020). In legal applications, however, retrieval alone is insufficient: even with the correct law in context, the model can still misquote a fragment, mix up article numbers, or fabricate a case reference. This motivates a *post-generation verification* layer that treats citations as structured objects and validates them against authoritative sources.

We present EscavAI, a production pipeline designed for this setting, and report its behavior over 184,895 audited answers. Our contributions are:

1. A **Reference-Audited RAG** architecture that couples domain-tuned hybrid retrieval with post-generation verification of every extracted legal reference, including existence checks at fragment granularity and fidelity-driven rewriting.
2. A practical **reference extraction and parsing stack** for Brazilian legal text, with taggers for legislation and jurisprudence mentions and a canonical schema for hierarchical fragment paths.
3. A **large-scale production evaluation** over

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 30–34

43,175 references that reveals a pronounced resolution gap between legislation (81.7%) and jurisprudence (47.1%), and shows that fidelity verification catches semantic errors in 6.5% of checked answers.

2 Domain Challenges in Brazilian Law

Two characteristics of the Brazilian legal domain make hallucination mitigation particularly challenging and motivate our design.

Citation complexity. A single provision can appear in many surface forms (e.g., *"Art. 5º, inciso LVII, da CF/88"* or *"artigo quinto, inciso 57 da Constituição Federal"*). References may contain enumerations ("§§ 2º, 3º, 6º e 8º do art. 11"), ranges, and nested paths. A verifier must map these variants to a canonical reference and resolve the cited fragment precisely.

Jurisdictional and instance constraints. For jurisprudence, correct identification depends on court, instance, decision type (*acórdão*, *súmula*, *recurso*), and process identifier formats that vary by tribunal. Models can generate syntactically valid but non-existent case numbers, or cite a valid number under the wrong court.

3 System Architecture

The system follows a four-stage pipeline: **Retrieve** → **Generate** → **Audit** → **Deliver** (Figure 1).

3.1 Stage 1: Hybrid Retrieval

Retrieval is implemented as a multi-phase funnel on a distributed search engine, combining lexical, semantic, and metadata signals. The knowledge base includes jurisprudence and editorial content (news, commentaries) in a vector/lexical index, while legislation and its fragments are served from a dedicated structured store.

Lexical and metadata retrieval. The first phase uses BM25 and field-aware lexical features (nativeRank, fieldMatch) with Portuguese stemming and tuned field weights. Concise legal summaries (*ementa*) receive substantially higher weight than full texts (*inteiro teor*) to limit noise from long documents. Metadata fields (tribunal, state, decision type, date) support fast filtering over millions of documents.

Dense retrieval. We index 768-dimensional embeddings produced by a domain-specific bi-encoder trained on millions of Brazilian legal document pairs using MultipleNegativesRankingLoss with hard negatives mined from top BM25 results. This stage increases recall for paraphrases and synonymy (*"dispensa imotivada"* vs. *"demissão sem justa causa"*), helping retrieval when the same legal idea is expressed with different wording.

Cross-encoder reranking. A quantized ONNX cross-encoder (Nogueira and Cho, 2019) is applied as a global phase on the top candidates. Documents below a confidence threshold of 0.2 are pruned; when no document exceeds this threshold, the system triggers a fallback response.

Query expansion. In the current deployed configuration, each user query is expanded into 7 reformulated sub-queries via function calling. Sub-queries execute in parallel across multiple collections (jurisprudence, *súmulas*, news, commented content), improving coverage of terminological variation.

Domain-specific calibration. Two practical issues required domain-specific adjustments: (i) a non-trivial fraction of jurisprudence entries lack an *ementa* field, requiring dynamic field-weight rebalancing to avoid burying relevant items; and (ii) very short document types (*súmulas*) can dominate lexical scoring due to their density, so we apply type-aware score calibration.

3.2 Stage 2: Grounded Generation

Retrieved documents are formatted into a context window with explicit source identifiers. The generation prompt enforces three constraints: (i) answer using only the retrieved context; (ii) use an explicit Brazilian legal citation style when mentioning authorities; and (iii) refuse when the answer cannot be supported by the available context.

Context budgeting. The system allocates a configurable token budget across entity types. A typical configuration assigns weight 0.50 to case law, 0.40 to commented jurisprudence and legal news, and 0.10 to *súmulas* and commented legislation. Unused budget from one type is redistributed to others, maintaining a total window of 13k to 24k tokens depending on query complexity and active filters.

3.3 Stage 3: Reference Auditing

The core novelty is a post-generation audit pipeline designed to prevent silent citation hallucinations. The audit operates on the *draft answer* and produces: (i) a structured list of audited references with per-reference status; and (ii) a marked-up answer where each reference is linked to the resolved authority when available.

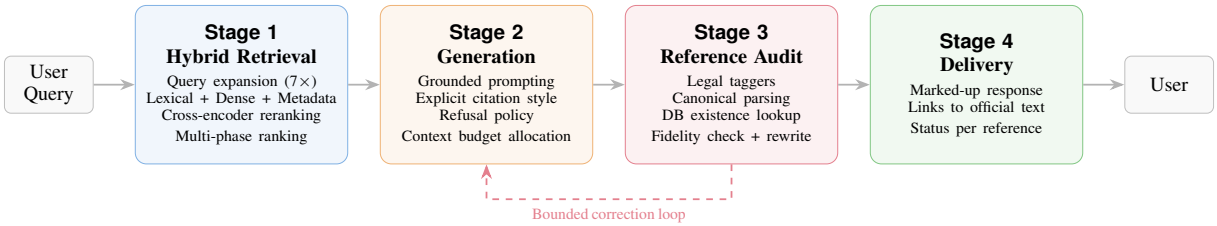

Figure 1: Pipeline overview. The audit stage can trigger a bounded rewrite loop with explicit error context and the retrieved official text for disputed references.

Step 1: Reference extraction. Two specialized taggers identify references in the generated text:

- A **Legislation Tagger** that recognizes mentions of laws, codes, constitutions, decrees, and their fragments. It handles significant surface variation, mapping *"Lei nº 8.666/93"*, *"Lei de Licitações"*, and *"L. 8666"* to the same canonical form.
- A **Jurisprudence Tagger** that identifies court decisions, *súmulas*, and temas, extracting court, decision type, and identifier.

Both taggers model the problem as a sequence labeling task trained on curated data constructed through a produce-consume pipeline: an LLM performs initial annotation, and human annotators validate and correct the output. The taggers are versioned and served via an internal microservice.

Step 2: Canonical parsing. Each tagged span is parsed into a structured object. For legislation, this includes type, number, year, and a fragment path (article → paragraph → inciso → alínea). For jurisprudence, the object includes court, decision type, and case identifier.

Step 3: Existence and fragment resolution. Each canonical object is resolved against authoritative stores: (i) a relational legislation database with fragment-level entries and temporal metadata; and (ii) a search index of case law covering multiple court levels. The audit assigns per-reference status: FOUND, PARTIALLY_FOUND (law exists but a specific fragment is missing), or NOT_FOUND. For legislation, status is also computed per individual fragment.

Step 4: Fidelity verification. When a reference is FOUND, the system compares what the draft answer claims about the authority against the retrieved official text. If the claim is unsupported, the system triggers a targeted rewrite prompt conditioned on the official excerpt, and re-validates the updated answer before delivery.

Table 1: Production telemetry summary at the reference level. RR = fully resolved references.

Metric	Legislation	Jurisprudence	Overall
References	38,914	4,261	43,175
RR%	81.7	47.1	78.2
Unresolved %	12.9	52.9	16.8

3.4 Stage 4: Delivery

When Step 4 triggers a correction, the rewrite loop is bounded (up to 2 iterations) to control latency. The final response is then returned with inline markup tags carrying reference identifiers and status values, plus a structured JSON audit report. The client UI can render links to official texts and highlight unresolved references visually.

4 Evaluation: Production Telemetry

We evaluate the system through automated reference-level audit signals on all audited production answers from launch to February 2026. The dataset contains **184,895 answers** and **43,175 extracted legal references** (38,914 legislation; 4,261 jurisprudence). Our primary goal is to quantify how often generated citations can be resolved to authoritative sources and how often fidelity checks require correction. Table 1 summarizes the reference-level results.

Across all audited answers, 22,595 (12.2%) include at least one explicit legislation or jurisprudence reference extracted by the audit pipeline. These answers account for the 43,175 references summarized in Table 1. Of those answers, 14,860 were fidelity-checked, and 961 were rewritten (6.5% of fidelity-checked answers). The resolution rates in Table 1 are reported at the reference level, whereas the rewrite rate is reported at the answer level.

Two findings are central. First, there is a large gap between legislation and jurisprudence resolution (81.7% vs. 47.1%), indicating that jurisprudence indexing/normalization remains the main bottleneck. Second, fidelity verification is necessary even when references resolve: 6.5% of checked

answers required rewriting, preventing incorrect legal interpretations from reaching end users. Operationally, the audit turns unresolved references into explicit warnings, shifting the failure mode from silent hallucination to transparent uncertainty.

4.1 Qualitative Error Analysis

To complement aggregate telemetry, we manually reviewed representative audit traces and observed four recurrent error classes:

1. **Jurisprudence coverage gaps:** the dominant source of NOT_FOUND case-law references, especially for recently published decisions across many courts.
2. **Long-tail legislation:** unresolved references to municipal, historical, or sparsely indexed norms.
3. **Fragment-level mismatches:** law-level matches where the cited paragraph/item does not resolve, captured as PARTIALLY_FOUND.
4. **Fidelity errors:** incorrect legal claims attached to real authorities, detected and corrected in the rewrite loop.

This analysis reinforces the value of surfacing unresolved references explicitly, rather than silently dropping them. It also points to the main priorities for future iterations: broader jurisprudence coverage and improved fragment-level resolution.

Representative production examples illustrate the behavior:

- *"Súmula 439 do STJ"* → FOUND (correct tribunal and identifier).
- *"§ 4º do art. 3º da Lei 13.105/2015"* → PARTIALLY_FOUND (law resolved, fragment unresolved).
- *"art. 4º da Constituição de 1946"* → NOT_FOUND (outside current coverage).

5 Deployment and Industry Impact

The pipeline is deployed in production at Escavador, a Brazilian legal information platform serving over 14 million monthly visitors. EscavAI supports multiple features: process summarization, process Q&A, general legal Q&A, movement explanation, document drafting, document review, and jurisprudence search, and currently serves over 60,000 monthly active users on AI-powered capabilities. From a compliance perspective, the system operates under Brazil's LGPD with NER-based PII masking. Audit traces provide a governance layer by recording cited authorities, resolution status, and supporting official excerpts.

6 Related Work

RAG is the dominant grounding strategy for LLM systems (Lewis et al., 2020), and strong retrieval stacks often combine hybrid retrieval with reranking (Karpukhin et al., 2020; Nogueira and Cho, 2019). In Brazilian Portuguese, prior legal-NLP resources and language models provide useful foundations for domain adaptation (Luz de Araujo et al., 2018; Souza et al., 2020; Polo et al., 2021; de Mello et al., 2024). For legal assistants, hallucinated authorities remain a key risk (Magesh et al., 2025). Corrective architectures such as Self-RAG and CRAG (Asai et al., 2024; Yan et al., 2024) motivate our domain-specific design choice: explicit post-generation auditing with fragment-level legal resolution and fidelity checking against official sources.

7 Conclusion

We described a production anti-hallucination pipeline for Brazilian Portuguese legal assistants, combining hybrid retrieval with a post-generation Reference Audit layer. Production telemetry over 184,895 answers and 43,175 references reveals that legislation references resolve at 81.7%, while jurisprudence references resolve at only 47.1%, identifying jurisprudence verification as the primary open challenge. Fidelity verification catches semantic errors in 6.5% of checked answers and triggers targeted rewrites before delivery.

The system shifts failures from silent hallucination to explicit uncertainty by flagging unresolved references. Future work includes temporal validity tracking, broader jurisprudence coverage, and improved handling of fragment-level mismatches.

References

A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proc. of ICLR*, 2024.

V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih. Dense passage retrieval for open-domain question answering. In *Proc. of EMNLP*, pages 6769–6781, 2020.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. of NeurIPS*, volume 33, pages 9459–9474, 2020.

P. H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo. LeNER-Br: A dataset for named entity recognition in Brazilian legal text. In *Proc. of PROPOR*, pages 313–323, 2018.

V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho. Hallucination-free? Assessing the reliability

of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(1):216–242, 2025.

G. L. de Mello, M. Finger, F. Serras, M. de Mello Carpi, M. M. Jose, P. H. Domingues, and P. Cavalin. PeLLE: Encoder-based language models for Brazilian Portuguese based on open data. In *Proc. of PROPOR*, pages 255–265, 2024.

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

F. M. Polo, G. C. F. Mendonça, K. C. J. Parreira, L. Gianvechio, P. Cordeiro, J. B. Ferreira, L. M. P. de Lima, A. C. do Amaral Maia, and R. Vicente. LegalNLP: Natural language processing methods for the Brazilian legal language. In *Proc. of ENIAC*, pages 763–774, 2021.

F. Souza, R. Nogueira, and R. Lotufo. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proc. of BRACIS*, pages 403–417, 2020.

S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling. Corrective retrieval augmented generation. In *Proc. of ICLR*, 2024.