Semantic Similarity Is Not Legal Correctness: Evaluating RAG Systems in Brazilian Civil Procedure

Eryclis Silva[1] , **Madelyn Sanfilippo**[1] ,
[1]University of Illinois at Urbana-Champaign
{ersilva2, madelyns}@illinois.edu,

Abstract

Evaluating retrieval-augmented generation (RAG) systems in legal domains is challenging due to the nuanced nature of legal reasoning and scarcity of domain-specific benchmarks. We investigate semantic similarity and legal correctness in Brazilian legal question answering, developing a synthetic dataset of 3,012 evaluation instances from the Brazilian Civil Procedure Code spanning seven query types, validated through human expert assessment. Our evaluation framework combines BERTScore with domain-adapted LLM-as-Judge (GPT-4o-mini), validated against expert legal assessment. Analysis reveals 50.8% disagreement between semantic similarity and legal correctness, with correlation varying substantially by query type ($\rho = 0.464$ to $\rho = 0.732$). Explicit article citation emerges as the strongest quality predictor (Cohen's $d = 1.099$), with cited responses achieving 145% higher legal correctness scores. Concrete examples demonstrate bidirectional divergence: responses may achieve high semantic similarity yet cite incorrect articles, or provide legally perfect answers with minimal lexical overlap. Findings demonstrate that semantic metrics alone are insufficient for legal RAG evaluation, with disagreement patterns structured by query type.

1 Introduction

Retrieval-Augmented Generation (RAG) systems [Gao *et al.*, 2024; Huang and Huang, 2024] demonstrate significant potential to enhance Large Language Models (LLMs) by incorporating external knowledge sources. Applications in specialized domains such as law present unique challenges due to complex regulations, technical terminology, and the necessity of grounding in authoritative sources. LLMs face critical limitations in legal contexts: knowledge obsolescence and plausible yet inaccurate outputs [Ji *et al.*, 2023; Yang *et al.*, 2023; Kalai *et al.*, 2025; Huang *et al.*, 2025; Xu *et al.*, 2025; Li *et al.*, 2024].

While recent research on legal NLP and RAG has focused on English-language corpora and common law jurisdictions [Nigam *et al.*, 2025; Kalra *et al.*, 2024; Pipitone and Alami, 2024; Guha *et al.*, 2023a], civil law systems remain underexplored [Zhang *et al.*, 2023; Ioannou *et al.*, 2025], particularly in non-English languages [Junior *et al.*, 2025]. Beyond data scarcity, legal evaluation introduces fundamental complexity: legal correctness requires interpretation of statutory language, synthesis of multiple provisions, and faithful application of principles [Nguyen *et al.*, 2025; Kant *et al.*, 2025]. Traditional semantic similarity metrics may inadequately capture whether responses cite correct articles or apply legal reasoning accurately, motivating investigation into evaluation methodology for legal RAG systems.

We address three research questions: **RQ1:** How do semantic similarity metrics align with legal correctness assessments, and how does alignment vary across query types? **RQ2:** What systematic patterns characterize response quality? **RQ3:** What are the strengths and limitations of LLM-as-Judge for legal RAG evaluation?

We develop a synthetic question-answer dataset spanning seven query types, validated through multi-agent critique and human expert assessment. Our evaluation framework combines semantic metrics (BERTScore) [Zhang *et al.*, 2020] with domain-adapted LLM-as-Judge (GPT-4o-mini), validated against expert legal assessment. Systematic analysis across 3,012 instances reveals 50.8% disagreement between semantic similarity and legal correctness, with correlation varying substantially by query type. Explicit article citation emerges as the strongest quality predictor, and concrete examples demonstrate bidirectional divergence between metrics. Findings establish that legal RAG evaluation requires multi-dimensional assessment combining complementary metrics.

2 Related Work

Legal NLP and Benchmarks. Legal NLP research has predominantly focused on English-language corpora and common law jurisdictions [Katz *et al.*, 2023; Chalkidis *et al.*, 2020], with benchmarks such as LegalBench [Guha *et al.*, 2023b] and COLIEE [Kim *et al.*, 2022] evaluating case outcome prediction and legal entailment in case law systems. Multilingual legal NLP remains limited [Ioannou *et al.*, 2025], with Portuguese resources particularly scarce. Civil law systems, relying on codified statutes rather than case precedent, present distinct challenges yet remain underexplored. Our work addresses this gap by focusing on Brazilian

civil procedure law, contributing evaluation methodology for statute-based legal systems.

RAG Evaluation Metrics. RAG evaluation frameworks [Gao *et al.*, 2024; Huang *et al.*, 2024; Yu *et al.*, 2024] typically employ general-domain metrics such as BLEU, ROUGE, and BERTScore [Lin, 2004; Papineni *et al.*, 2002; Zhang *et al.*, 2020], which assess lexical overlap and semantic similarity. Recent work explores automated evaluation [Es *et al.*, 2024; Saad-Falcon *et al.*, 2024], but frameworks predominantly target general domains without adaptation for specialized fields requiring domain-specific correctness. Our work demonstrates that semantic metrics systematically diverge from legal correctness (50.8% disagreement), establishing the necessity of multi-dimensional evaluation in legal domains.

LLM-as-Judge in Specialized Domains. LLM-as-Judge approaches [Zheng *et al.*, 2023] demonstrate correlation with human judgments in general tasks [Liu *et al.*, 2023], but effectiveness in specialized domains requires validation. Legal evaluation presents unique challenges: correctness depends on accurate statutory citation, faithful application of principles, and synthesis of multiple provisions—dimensions orthogonal to surface-level fluency. We validate LLM-as-Judge against expert legal assessment (Spearman $\rho = 0.96$), demonstrating viability for legal evaluation when calibrated through domain-specific prompting and reference-guided frameworks.

Our work systematically analyzes metric divergence in legal RAG evaluation, revealing that correlation between semantic similarity and legal correctness varies substantially by query type ($\rho = 0.464$ to $\rho = 0.732$), with explicit citation as the strongest quality predictor (Cohen's $d = 1.099$). Unlike previous work assuming metric alignment, we establish evaluation methodology itself as a critical challenge in specialized domains.

3 Methodology

To investigate systematic divergence between semantic similarity and legal correctness in RAG evaluation, we employ a controlled experimental design with three core components. First, we construct a synthetic dataset of 502 question-answer pairs from Brazilian civil procedure law, spanning seven query types that capture diverse legal reasoning patterns (simple factual retrieval, multi-step reasoning, comparative analysis, hypothetical application). Second, we implement six RAG configurations combining lexical, semantic, and hybrid retrieval strategies with varying context sizes, enabling comparative analysis across 3,012 evaluation instances. Third, we assess each generated response using complementary metrics—BERTScore for semantic similarity and GPT-4o-mini LLM-as-Judge for legal correctness—with both automated and human expert validation to ensure measurement reliability. This design enables us to systematically quantify metric divergence (RQ1), characterize failure patterns (RQ2), and evaluate domain-adapted assessment approaches (RQ3).

3.1 Synthetic Dataset Generation

LLM-generated synthetic datasets enable efficient evaluation data creation when manual annotation is costly [Long *et al.*, 2024; Liu *et al.*, 2024; Li *et al.*, 2023]. Our generation pipeline comprises context construction, question-answer generation, multi-agent validation, and human expert assessment.

Context Construction
We sourced contexts from the Brazilian Civil Procedure Code (*Código de Processo Civil*, CPC), the primary statute governing civil litigation procedures in Brazil. The CPC was segmented respecting its legal article structure, from which we randomly sampled 100 articles (approximately 10% of the corpus) as base units. To create contexts suitable for complex queries requiring broader semantic coverage, each base article was enhanced with its top-5 most semantically related articles[1] based on cross-references, conceptual proximity, and thematic coherence within the legal framework.

Question-Answer Pair Generation
Using Mistral 7B [Jiang *et al.*, 2023], we generated 100 candidate question-answer pairs per query type. We designed seven query types derived from established legal reasoning categories in legal education literature [Nguyen *et al.*, 2025]: *simple* (direct factual queries), *complex* (multi-concept synthesis), *adversarial* (queries with misleading elements), *situational* (user-contextualized scenarios), *dual* (two subquestions), *comparative* (comparing legal frameworks), and *hypothetical* (applying law to novel cases). This taxonomy captures diverse cognitive demands: simple queries test retrieval accuracy, comparative queries require synthesis across provisions, and hypothetical queries demand legal reasoning application. Generation yielded 700 total candidate pairs.

Multi-Agent Validation
To ensure dataset quality, we employed three critique agents implemented with Llama 3 8B [Grattafiori *et al.*, 2024], each assessing a distinct dimension [Asai *et al.*, 2023]. The *Groundedness Agent* evaluates whether context sufficiently supports answering the question (1-5 scale, where 1 indicates insufficient support and 5 indicates complete answerability). The *Relevance Agent* assesses whether the question addresses meaningful legal problem-solving. The *StandAlone Agent* determines whether the question is self-contained or requires external context. We retained only pairs scoring ≥ 3 across all three agents, reducing 700 candidates to 502 high-quality pairs (71.7% retention).[2] Full critique agent prompts are provided in Appendix A.

Human Expert Validation
We conducted expert validation to assess synthetic dataset quality and calibrate automated evaluation metrics. A group of Brazilian attorneys with civil litigation experience performed two validation tasks across three independent annotation sessions.

[1] Semantic similarity computed via cosine distance of embeddings from `paraphrase-multilingual-mpnet-base-v2` [Reimers and Gurevych, 2019].

[2] Threshold ablation showed that ≥ 2 retained 89% (625 pairs) with lower quality, while ≥ 4 retained 54% (380 pairs) with excessive filtering. Threshold ≥ 3 balanced quality and diversity, subsequently validated by expert assessment.

Type	Description	Count
Simple	Direct queries from legal text	74
Complex	Multi-concept or rephrased queries	78
Adversarial	Queries with misleading information	51
Situational	User-contextualized queries	70
Dual	Two subquestions in one query	74
Comparative	Comparing legal concepts/frameworks	93
Hypothetical	Applying law to hypothetical scenarios	62
Total dataset size		**502**
Total instances (502 × 6 methods)		3,012

Table 1: Distribution of query types in the final validated dataset.

Query Realism Assessment. The experts evaluated 105 synthetic queries (15 per type) on legal realism using a 5-point scale. Results showed that 91.1% of queries achieved high realism scores ($\geq$4), confirming ecological validity across all query types. Complete statistics are provided in Appendix B.

LLM-Judge Calibration. The experts scored 50 query-answer pairs on legal correctness. Both experts and LLM-Judge performed reference-guided evaluation: each received the query, generated answer, and gold-standard reference answer, then assessed legal correctness by comparing the generated response against the reference. This comparative framework—standard in legal assessment—provides explicit evaluation criteria. Experts judgments demonstrated strong correlation with GPT-4o-mini LLM-as-Judge (consensus $\rho = 0.96$, $p < 0.001$; ICC=0.95), validating it as a reliable proxy for expert legal assessment. Detailed correlation analysis and annotation protocols are provided in Appendix B.

Table 1 presents the final distribution of query types across the validated dataset.

3.2 RAG System Implementation

We evaluate RAG performance across multiple retrieval strategies and context sizes to enable systematic comparison of metric behavior under different system configurations.

Retrieval Strategies

We implemented three retrieval approaches, each evaluated with two context sizes ($k \in \{3, 7\}$), yielding six configurations. *BM25* (parameters: $k_1 = 1.2$, $b = 0.75$) serves as our lexical baseline, ranking documents by term overlap [Robertson and Zaragoza, 2009]. *Dense retrieval* encodes documents into 768-dimensional semantic vectors using sentence embeddings,[3] indexed using FAISS for cosine similarity computation [Douze *et al.*, 2025]. *Hybrid retrieval* combines BM25 and dense scores through linear interpolation: for each document d, we compute

$$\text{score}(d) = \alpha \cdot \frac{1}{\text{dist}_{\text{dense}}(d) + \epsilon} + (1 - \alpha) \cdot \text{score}_{\text{BM25}}(d) \quad (1)$$

where $\epsilon = 10^{-6}$ prevents division by zero and $\alpha = 0.5$ equally weights semantic and lexical signals, avoiding bias toward either retrieval paradigm.[4] Context sizes $k = 3$ and $k = 7$ represent minimal and extended context scenarios, respectively, enabling analysis of how context volume affects metric alignment.

Generation Component

Retrieved documents are provided as context to Llama 3 8B Instruct [Grattafiori *et al.*, 2024] for answer generation. The system prompt instructs the model to ground responses in the provided legal context and cite specific CPC articles when making legal references, though explicit citation is not mandated. This design reflects realistic deployment scenarios where models are encouraged but not forced to cite sources. The complete generation prompt is provided in Appendix A.

3.3 Evaluation Framework

We employ a dual-metric evaluation approach to assess both surface-level semantic coherence and substantive legal correctness, enabling systematic analysis of their alignment.

Semantic Similarity Metrics

We use BERTScore F1 [Zhang *et al.*, 2020] to measure semantic similarity between generated and reference answers. BERTScore computes token-level similarity using contextualized embeddings, capturing semantic alignment beyond lexical overlap. We selected BERTScore over traditional metrics (BLEU, ROUGE) due to its superior performance in capturing contextual similarity and widespread adoption in text generation evaluation [Gao *et al.*, 2025]. BERTScore scores range from 0 to 1, with higher values indicating greater semantic similarity to the reference answer.

Legal Correctness Assessment

We employ GPT-4o-mini as an LLM-as-Judge [Zheng *et al.*, 2023] to assess legal correctness, a dimension where lexical metrics systematically fail to capture domain-specific reasoning [Oliva *et al.*, 2025; Hu and Zhou, 2024]. The judge operates in a reference-guided configuration: for each evaluation instance, it receives the query, generated answer, and gold-standard reference answer, then scores legal correctness on a 1-5 scale by comparing the generated response against the reference. This comparative evaluation framework—standard in legal assessment—provides explicit criteria for judgment: correct article citations, accurate legal interpretation, and completeness of reasoning.

The LLM-Judge prompt (Appendix A) emphasizes assessment of substantive legal grounding rather than surface fluency. Critically, our human validation study (Section 3.1.4) demonstrated strong correlation between GPT-4o-mini and expert legal judgment (consensus $\rho = 0.96$, $p < 0.001$; ICC=0.95), validating its use as a reliable proxy for expert assessment. LLM-Judge scores are normalized to [0,1] for comparison with BERTScore.

Statistical Analysis

We quantify metric divergence as the absolute difference between normalized scores: divergence $= |\text{BERT}_{\text{norm}} - \text{LLM}_{\text{norm}}|$, where values near 0 indicate alignment and values

[3] We use the same embedding model as context construction (`paraphrase-multilingual-mpnet-base-v2`) for consistency.

[4] Preliminary analysis showed hybrid performance remained stable across $\alpha \in [0.3, 0.7]$, confirming robustness of equal weighting.

near 1 indicate maximum disagreement. Correlation analysis employs Spearman's rank correlation coefficient (ρ) to assess monotonic relationships between metrics, chosen for its robustness to non-linear associations. Statistical significance is tested using permutation tests with $p < 0.05$ threshold.

4 Results

4.1 Evaluation Landscape

Evaluation across 3,012 question-answer instances reveals systematic patterns between semantic similarity (BERTScore) and legal correctness (LLM-as-Judge). Our analysis employs GPT-4o-mini as the LLM-Judge, validated against expert legal assessment with Spearman correlation of $\rho = 0.96$ ($p < 0.001$) and intraclass correlation coefficient of ICC=0.95 (Section 3.1.4). Overall performance shows BERTScore mean of 0.676 ± 0.100 and LLM-Judge mean of 0.415 ± 0.322, with metric divergence of 0.305 ± 0.230. The threefold difference in standard deviation (0.100 vs 0.322) reflects that semantic similarity assessments are far more consistent than legal correctness evaluations, suggesting the latter captures nuanced domain-specific quality dimensions.

Tables 2 and 3 present granular scores for each retrieval strategy across query types. Three key observations emerge: (1) BERTScore values are consistently higher and more uniform across methods (0.63–0.71) compared to LLM-Judge scores (0.13–0.66), indicating that semantic coherence is substantially easier to achieve than legal correctness; (2) Comparative queries exhibit the lowest performance on both metrics, particularly for LLM-Judge (0.13–0.29 vs 0.45–0.66 for other types), suggesting fundamental challenges in comparative legal reasoning that persist across retrieval strategies; (3) Dense retrieval with $k = 7$ modestly outperforms BM25 and hybrid methods on LLM-Judge for most query types, though no method achieves consistently high legal correctness scores.

4.2 Metric Divergence Patterns

Agreement Analysis

To investigate the alignment between semantic and legal quality, we categorized both BERTScore and LLM-Judge into three discrete levels (Low < 0.33, Medium 0.33–0.66, High > 0.66) and analyzed their cross-distribution. Table 4 presents the resulting confusion matrix, revealing substantial divergence between the two metrics.

Only 49.2% of responses (1,483 cases) fall along the main diagonal, indicating agreement between BERTScore and LLM-Judge. The remaining 50.8% (1,529 cases) exhibit disagreement, with 8.9% (268 cases) classified as extreme disagreement—cases where one metric assigns Low while the other assigns High. This near-parity between agreement and disagreement demonstrates that semantic similarity and legal correctness measure fundamentally orthogonal quality dimensions in the legal domain.

The confusion patterns reveal systematic asymmetry. BERTScore frequently overestimates legal quality: 22.0% of responses (664 cases) receive Medium BERTScore but Low LLM-Judge, while only 6.8% (206 cases) show High BERTScore with Low LLM-Judge. Conversely, the reverse pattern—high legal correctness with low semantic similarity—is rare, occurring in only 2.1% of cases (62). This asymmetry indicates that achieving semantic coherence in legal responses is necessary but insufficient for legal correctness, as fluent text may cite incorrect articles, misapply legal provisions, or provide superficially plausible but substantively wrong answers.

Agreement is strongest in the High-High category (24.3%, 732 cases), suggesting that responses demonstrating both semantic fluency and legal correctness represent a minority of outputs. The Low-Low category accounts for 13.6% (410 cases), capturing responses that fail on both dimensions—typically "don't know" refusals or severely truncated answers. Medium-Medium agreement is the weakest (11.3%, 341 cases), reflecting that moderate-quality responses are particularly prone to divergence between semantic and legal assessment.

Correlation by Query Type

Table 5 presents Spearman correlation coefficients between BERTScore and LLM-Judge across query types. All correlations are statistically significant ($p < 0.001$), indicating that semantic quality and legal correctness demonstrate positive association across all query categories. However, correlation magnitude varies substantially, ranging from $\rho = 0.464$ (situational) to $\rho = 0.732$ (hypothetical), revealing that query type fundamentally structures the alignment between semantic and legal assessment.

Hypothetical queries exhibit the strongest correlation ($\rho = 0.732$), suggesting that speculative legal scenarios elicit responses where semantic fluency reliably indicates legal correctness. Adversarial queries also demonstrate strong correlation ($\rho = 0.637$), indicating that deliberately misleading prompts tend to produce clear success or failure patterns. Situational queries show the weakest correlation ($\rho = 0.464$), despite achieving moderate LLM-Judge scores (mean: 0.479), suggesting that contextualized legal queries produce responses with variable alignment between semantic coherence and legal accuracy. Simple and dual queries fall in the intermediate range ($\rho = 0.502$ for both).

Comparative queries demonstrate moderate correlation ($\rho = 0.633$) despite lowest absolute performance (LLM mean: 0.185). This reflects that 72.2% are refusals or partial responses: BERTScore captures lexical overlap even in unsuccessful answers, while LLM-Judge strictly penalizes refusal markers, producing consistent divergence patterns. Section 4.3 examines comparative query failure modes in detail.

4.3 Quality Indicators and Response Patterns

Citation as Quality Predictor

Explicit citation of legal articles emerges as the strongest quality indicator in our dataset. Among 3,012 responses, 63.6% (1,917 cases) contain explicit article references, achieving mean LLM-Judge score of 0.529 ± 0.255 compared to 0.216 ± 0.332 for responses without citations. This represents a 145% performance increase, with Cohen's $d = 1.099$ indicating a large effect size. BERTScore also shows positive association with citation (0.708 vs 0.621), though the effect

Method	Comp.	Complex	Advers.	Dual	Hypoth.	Simple	Situat.
BM25 ($k = 3$)	0.631	0.677	0.670	0.687	0.659	0.683	0.685
BM25 ($k = 7$)	0.628	0.684	0.688	0.698	**0.687**	0.685	0.694
Dense Ret. ($k = 3$)	0.636	0.684	0.648	0.701	0.650	0.691	0.689
Dense Ret. ($k = 7$)	**0.654**	**0.712**	0.661	**0.705**	0.686	**0.696**	**0.702**
Hybrid ($k = 3$)	0.631	0.677	0.672	0.689	0.660	0.687	0.684
Hybrid ($k = 7$)	0.630	0.689	**0.695**	0.699	0.682	0.686	0.695

Table 2: BERTScore F1 by retrieval strategy and query type. Dense retrieval with $k = 7$ achieves highest scores across most query types, with Comparative queries consistently showing lowest performance.

Method	Comp.	Complex	Advers.	Dual	Hypoth.	Simple	Situat.
BM25 ($k = 3$)	0.132	0.474	0.451	0.334	0.427	0.480	0.407
BM25 ($k = 7$)	0.202	0.458	0.471	0.422	0.464	0.561	0.521
Dense Ret. ($k = 3$)	0.167	0.452	0.348	0.463	0.407	0.507	0.464
Dense Ret. ($k = 7$)	**0.285**	**0.522**	0.412	**0.510**	**0.468**	**0.578**	**0.550**
Hybrid ($k = 3$)	0.129	0.471	0.456	0.348	0.431	0.493	0.404
Hybrid ($k = 7$)	0.199	0.487	**0.500**	0.449	0.456	0.568	0.525

Table 3: LLM-as-Judge scores (GPT-4o-mini) by retrieval strategy and query type. Comparative queries exhibit substantially lower legal correctness (0.13–0.29) compared to other types, with Dense $k = 7$ consistently achieving the highest scores across most query types.

	LLM-Judge			
BERTScore	**Low**	**Medium**	**High**	**Total**
Low	**410** (13.6%)	42 (1.4%)	62 (2.1%)	514 (17.1%)
Medium	664 (22.0%)	**341** (11.3%)	187 (6.2%)	1,192 (39.6%)
High	206 (6.8%)	368 (12.2%)	**732** (24.3%)	1,306 (43.4%)
Total	1,280 (42.5%)	751 (24.9%)	981 (32.6%)	**3,012**

Table 4: Confusion matrix of BERTScore and LLM-Judge agreement. Metrics categorized as Low (< 0.33), Medium (0.33–0.66), or High (> 0.66). Diagonal entries (bold) indicate agreement (46.7%); off-diagonal entries indicate disagreement (53.3%). Cell shading intensity reflects frequency magnitude.

Query Type	# Cases	ρ	BERTScore		LLM-Judge	
			Mean	**Std**	**Mean**	**Std**
Situational	420	0.464	0.691	0.074	0.479	0.292
Simple	444	0.502	0.688	0.109	**0.531**	0.345
Dual	444	0.502	**0.696**	0.063	0.421	0.272
Complex	468	0.520	0.687	0.090	0.478	0.284
Comparative	558	0.633	0.635	0.095	0.185	0.265
Adversarial	306	0.637	0.672	0.131	0.440	0.334
Hypothetical	372	**0.732**	0.671	0.121	0.442	0.334

Table 5: Spearman correlation between BERTScore and LLM-Judge by query type. Hypothetical queries show strongest correlation ($\rho = 0.732$), while Situational queries show weakest ($\rho = 0.464$). All correlations significant at $p < 0.001$.

is substantially smaller, suggesting that citation primarily predicts legal correctness rather than semantic fluency.

Table 6 presents citation patterns and their impact across query types. Citation rates vary substantially, ranging from 42.7% for comparative queries—which exhibit highest failure rates—to 77.0% for dual queries. Despite this variation, the positive association between citation and quality is consistent across all query types. Hypothetical queries show the largest citation effect ($\Delta = 0.432$), where cited responses achieve nearly four times the score of uncited ones (0.590 vs 0.157). Simple queries exhibit the smallest effect ($\Delta = 0.129$), as straightforward legal questions may be adequately answered without explicit statutory grounding. Comparative queries, despite low citation rates, demonstrate substantial impact when citations are present ($\Delta = 0.343$), indicating that successful comparative reasoning critically depends on article grounding.

The consistent positive association between explicit citation and both semantic and legal quality metrics demonstrates that article references serve as reliable indicators of response grounding. This suggests potential for automated quality assessment through citation extraction, enabling filtering or reranking of RAG outputs without expensive LLM evaluation—a design implication discussed in Section 5.

Response Length Effects

Response length demonstrates moderate correlation with evaluation metrics and divergence patterns. Responses were categorized into four bins: Very Short (<50 words), Short (50–100), Medium (100–200), and Long (>200). The distribution is heavily skewed toward very short responses (67.6%, 2,036 cases), while long responses are rare (0.2%, 7 cases).

Both BERTScore and LLM-Judge increase with response length (Table 7), rising from 0.656 and 0.358 respectively for very short responses to 0.727 and 0.536 for long responses. Correspondingly, divergence decreases from 0.351 to 0.204, indicating that longer responses achieve better alignment between semantic and legal quality. Correlation analysis reveals differential metric sensitivity: BERTScore shows mod-

Query Type	Total (n)	Citation Rate	LLM-Judge Score With	Without	Δ
Hypothetical	372	65.9%	0.590	0.157	**+0.432**
Comparative	558	42.7%	0.382	0.039	+0.343
Adversarial	306	57.2%	0.577	0.256	+0.321
Situational	420	75.0%	0.546	0.276	+0.270
Dual	444	**77.0%**	0.480	0.223	+0.257
Complex	468	67.3%	0.557	0.314	+0.243
Simple	444	64.6%	0.577	0.447	+0.129

Table 6: Citation patterns and quality impact by query type. Citation Rate shows percentage of responses with explicit article references. LLM-Judge scores are reported for responses with and without citations, along with the performance gain (Δ). Query types are ordered by citation effect magnitude.

Length Category	n	BERT	LLM	Div.	Words
Very Short (<50)	2,036	0.656	0.358	0.351	26
Short (50–100)	785	0.715	0.516	0.221	70
Medium (100–200)	184	0.725	**0.606**	**0.151**	130
Long (>200)	7	**0.727**	0.536	0.204	218

Table 7: Evaluation metrics across response length categories. BERTScore and LLM-Judge improve with length, while divergence decreases, indicating better metric alignment in longer responses.

erate correlation with word length (Spearman $\rho = 0.447$, $p < 0.001$), while LLM-Judge exhibits weaker correlation ($\rho = 0.359$, $p < 0.001$). The strong negative correlation between LLM-Judge and divergence ($\rho = -0.720$, $p < 0.001$) indicates that legally correct responses achieve better metric agreement regardless of length.

Very short responses (mean: 26 words) lack detail for legal context, qualifications, and supporting rationale—elements essential for accurate legal analysis. For instance, "*The deadline is 15 days*" omits critical qualifications, while "*The deadline is 15 business days from service of process, excluding weekends and holidays, as established in Article 335*" captures legal precision and grounding. The decreasing divergence with length ($\Delta = 0.147$ from Very Short to Long) suggests that both semantic similarity and legal grounding converge when responses have adequate scope to express legal nuance, highlighting a fundamental challenge for RAG systems: balancing concision with the elaboration required for legally sound reasoning.

4.4 Concrete Examples of Metric Divergence

Analysis of cases exhibiting substantial metric disagreement reveals systematic patterns in how semantic similarity and legal correctness diverge. These examples demonstrate that divergence between BERTScore and LLM-Judge reflects fundamentally different quality dimensions: surface-level linguistic coherence versus substantive legal accuracy.

Scenario 1: Semantic Coherence Without Legal Correctness. Consider a complex query requesting information about courtroom recording procedures, document illustration, and evidentiary standards:[5]

> *"In a judicial proceeding, can rapid recording techniques such as shorthand or stenography be used to record court deliberations? Additionally, can the case file be illustrated with drawings, graphs, or photographs? What evidence is necessary for merit judgment and what can be dismissed by the judge?"*

The generated response achieved high BERTScore (0.802) by addressing all three sub-questions with fluent legal terminology, producing strong lexical overlap with the reference answer. However, it received low LLM-Judge score (0.250, divergence: 0.552) due to fundamentally incorrect article citations: Article 343 for recording procedures (correct: Article 210), Article 3º §2º of Article 2º for visual documentation (correct: Article 484), and Article 123 for evidentiary standards (correct: Article 370). These citations reference entirely different legal provisions addressing unrelated procedural matters. BERTScore rewards the presence of legal concepts ("shorthand," "photographs," "evidence") regardless of whether supporting citations are correct, while LLM-Judge correctly identifies that incorrect statutory references render the response legally unreliable despite semantic fluency.

Scenario 2: Legal Correctness Without Semantic Similarity. A simple query regarding expert fee proposal deadlines elicited a seven-word response:[6]

> *Query:* "What is the deadline for submitting the expert's fee proposal after appointment?"
> *Generated:* "5 (five) days. (Art. 465, § 2º)"
> *Reference:* "The expert must submit their fee proposal within 5 (five) days after appointment."

This response received low BERTScore (0.565) due to minimal lexical overlap—omitting "expert," "submit," "proposal," "fee," and "appointment"—while achieving perfect LLM-Judge score (1.000, divergence: 0.435) by providing the legally essential information: the specific deadline and authoritative source. The reference answer's additional context improves readability but adds no legal substance. This asymmetry reveals BERTScore's sensitivity to verbosity: semantic metrics reward lexical redundancy, while legal correctness depends on citation accuracy and factual precision, not elaboration.

These cases validate multi-dimensional evaluation necessity in legal RAG. Semantic similarity captures linguistic properties that may coexist with legal incorrectness or be absent despite legal correctness. Legal correctness depends on dimensions orthogonal to surface form: accurate statutory citation, correct interpretation, and faithful application of principles. Evaluation frameworks relying solely on semantic metrics risk systematically misclassifying response quality in both directions.

5 Discussion

Our investigation of 3,012 question-answer instances reveals that semantic similarity and legal correctness represent fun-

[5]Query ID: new_results_bm25_k_7_129. Full text in Appendix C. Translated from Portuguese.

[6]Query ID: new_results_hybrid_linear_k_3_350. Translated from Portuguese.

damentally orthogonal quality dimensions, with systematic misalignment structured by query type and response characteristics.

5.1 Addressing the Research Questions

RQ1: Metric Alignment Across Query Types. Analysis reveals 49.2% agreement between BERTScore and LLM-Judge, with 50.8% exhibiting disagreement. Critically, alignment varies substantially by query type: Spearman correlations range from $\rho = 0.464$ (situational) to $\rho = 0.732$ (hypothetical), demonstrating that query complexity fundamentally influences whether semantic coherence corresponds to legal correctness. This variation necessitates multi-metric evaluation frameworks and suggests query-adaptive strategies that weight metrics based on query type may improve assessment accuracy.

RQ2: Quality Patterns and Predictors. Explicit article citation emerged as the strongest quality predictor (Cohen's $d = 1.099$): responses with citations achieved 145% higher LLM-Judge scores (0.529 vs 0.216), with effects consistent across query types ($\Delta = 0.129$ to $\Delta = 0.432$). Response length showed moderate correlation with quality ($\rho = 0.359$), with divergence decreasing from 0.351 for very short responses to 0.204 for long responses. Concrete examples demonstrate bidirectional divergence: responses may achieve high BERTScore through fluent terminology yet cite incorrect articles (BERT=0.802, LLM=0.250), or provide legally perfect answers with minimal lexical overlap (BERT=0.565, LLM=1.000). Citation presence enables automatic quality assessment without expensive LLM evaluation.

RQ3: LLM-as-Judge Evaluation. GPT-4o-mini demonstrated strong correlation with expert assessment (Spearman $\rho = 0.96$, ICC=0.95), successfully capturing legal nuance that semantic metrics miss. The approach correctly identifies incorrect legal grounding despite semantic coherence, as demonstrated by responses citing wrong articles yet maintaining fluent language. Systematic sensitivities include strong preference for explicit citations (2.5× score difference) and correlation with response length ($\rho = 0.359$), suggesting LLM-as-Judge measures a specific conception of legal quality emphasizing citation and elaboration that aligns with practitioner judgments.

5.2 Implications for Legal RAG Evaluation

The 50.8% disagreement rate establishes that semantic metrics alone systematically misclassify quality in both directions—fluent responses may contain legal errors while concise responses may be legally perfect. Evaluation frameworks must combine complementary metrics capturing distinct dimensions: semantic metrics for linguistic quality, domain-specific assessment for substantive correctness. The substantial correlation variation ($\rho = 0.464$ to $\rho = 0.732$) across query types suggests query-adaptive evaluation strategies may outperform uniform approaches. The large citation effect (Cohen's $d = 1.099$) demonstrates that automatic signal extraction enables practical two-stage evaluation: lightweight filtering via citation detection followed by selective detailed assessment for ambiguous cases.

5.3 Limitations

Our evaluation employs synthetically generated question-answer pairs, which ensures controlled assessment but may not capture linguistic variability in practitioner queries. LLM-Judge validation involved three legal experts across limited sessions; broader validation would strengthen generalizability. Analysis focuses on Brazilian civil procedure law; patterns may differ in common law systems or other civil law jurisdictions. The reference-guided evaluation framework constrains LLM-Judge to comparative rather than absolute assessment, potentially missing valid alternative interpretations. More sophisticated retrieval methods may alter specific patterns, though the fundamental orthogonality between semantic similarity and legal correctness likely persists across approaches.

6 Conclusion

We investigated metric divergence in Brazilian legal RAG evaluation through systematic analysis of 3,012 question-answer instances. Our findings demonstrate that semantic similarity and legal correctness represent fundamentally orthogonal quality dimensions, with 50.8% disagreement between BERTScore and LLM-as-Judge. Correlation between metrics varies substantially by query type (Spearman $\rho = 0.464$ to $\rho = 0.732$), establishing that query complexity fundamentally structures the relationship between semantic coherence and legal accuracy. Explicit article citation emerges as the strongest quality predictor (Cohen's $d = 1.099$), with cited responses achieving 145% higher legal correctness scores, suggesting automatic citation extraction as a viable quality signal for practical deployment.

Concrete examples illustrate bidirectional divergence: responses may achieve high semantic similarity through fluent terminology yet cite incorrect articles, or provide legally perfect answers with minimal lexical overlap. LLM-as-Judge, validated against expert legal assessment (Spearman $\rho = 0.96$), successfully captures legal nuance that traditional metrics miss, though systematic sensitivities to citation format and verbosity require consideration in deployment.

These results establish that evaluation frameworks for legal RAG systems require multi-dimensional assessment combining complementary metrics. Query-adaptive strategies that weight semantic and legal metrics based on query type may improve evaluation accuracy. Our methodology—synthetic data generation with multi-agent validation, human expert calibration, and systematic metric divergence analysis—provides a foundation for evaluation research in specialized domains beyond law.

References

[Asai *et al.*, 2023] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511, 2023.

[Chalkidis *et al.*, 2020] Ilias Chalkidis, Manos Fergadiotis, et al. LEGAL-BERT: The muppets straight out of law school. In Trevor Cohn, Yulan He, and Yang Liu, editors,

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2898–2904, Online, November 2020. Association for Computational Linguistics.

[Douze *et al.*, 2025] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025.

[Es *et al.*, 2024] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In Nikolaos Aletras and Orphee De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[Gao *et al.*, 2024] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[Gao *et al.*, 2025] Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687, June 2025.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models, 2024.

[Guha *et al.*, 2023a] Neel Guha, Julian Nyarko, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

[Guha *et al.*, 2023b] Neel Guha, Julian Nyarko, et al. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[Hu and Zhou, 2024] Taojun Hu and Xiao-Hua Zhou. Unveiling llm evaluation focused on metrics: Challenges and solutions, 2024.

[Huang and Huang, 2024] Yizheng Huang and Jimmy Huang. A survey on retrieval-augmented text generation for large language models, 2024.

[Huang *et al.*, 2024] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, November 2024.

[Huang *et al.*, 2025] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025.

[Ioannou *et al.*, 2025] Antreas Ioannou, Andreas Shiamishis, Nora Hollenstein, and Nezihe Merve Gürel. Evaluating the limits of large language models in multilingual legal reasoning, 2025.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, et al. Mistral 7b, 2023.

[Junior *et al.*, 2025] Roseval Malaquias Junior, Ramon Pires, et al. Juru: Legal brazilian large language model from reputable sources, 2025.

[Kalai *et al.*, 2025] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025.

[Kalra *et al.*, 2024] Rishi Kalra, Zekun Wu, et al. HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications. In Sachin Kumar, Vidhisha Balachandran, Chan Young Park, Weijia Shi, Shirley Anugrah Hayati, Yulia Tsvetkov, Noah Smith, Hannaneh Hajishirzi, Dongyeop Kang, and David Jurgens, editors, *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 237–256, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Kant *et al.*, 2025] Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. Towards robust legal reasoning: Harnessing logical llms in law, 2025.

[Katz *et al.*, 2023] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II. Natural language processing in the legal domain, 2023.

[Kim *et al.*, 2022] Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. Coliee 2022 summary: Methods fornbsp;legal document retrieval andnbsp;entailment. In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12–17, 2022, Revised Selected Papers*, page 51–67, Berlin, Heidelberg, 2022. Springer-Verlag.

[Li *et al.*, 2023] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore, December 2023. Association for Computational Linguistics.

[Li *et al.*, 2024] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The

dawn after the dark: An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Liu *et al.*, 2023] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[Liu *et al.*, 2024] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data, 2024.

[Long *et al.*, 2024] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Nguyen *et al.*, 2025] Ha Thanh Nguyen, Wachara Fungwacharakorn, et al. Llms for legal reasoning: A unified framework and future perspectives. *Computer Law Security Review*, 58:106165, 2025.

[Nigam *et al.*, 2025] Shubham Kumar Nigam, Balaramamahanthi Deepak Patnaik, et al. Nyayarag: Realistic legal judgment prediction with rag under the indian common law system, 2025.

[Oliva *et al.*, 2025] Maria Paz Oliva, Adriana Correia, Ivan Vankov, and Viktor Botev. The illusion of a perfect metric: Why evaluating ai's words is harder than it looks, 2025.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[Pipitone and Alami, 2024] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain, 2024.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[Robertson and Zaragoza, 2009] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

[Saad-Falcon *et al.*, 2024] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[Xu *et al.*, 2025] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2025.

[Yang *et al.*, 2023] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. 2023.

[Yu *et al.*, 2024] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey, 2024.

[Zhang *et al.*, 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[Zhang *et al.*, 2023] Xiang Zhang, Senyu Li, et al. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore, December 2023. Association for Computational Linguistics.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.