Application of Retrieval-Augmented Generation (RAG) in the Automated Extraction of Legal Arguments in Labor Administrative Defenses

Rubens Cauan Figueredo Castro[1], **Regis Pires Magalhães**[1], **Luis Gustavo Coutinho do Rêgo**[2], **Alexandre Antonio Bruno da Silva**[3]

[1] Universidade Federal do Ceará (UFC)

[2]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

[3]Universidade Estadual do Ceará (UECE)

rubenscauanfc2021@alu.ufc.br, regismagalhaes@ufc.br

gustavo.coutinho@ifce.edu.br, alexandreantoniobruno@gmail.com

Abstract. *This paper investigates the use of Retrieval-Augmented Generation (RAG) for the automated extraction of legal arguments from statements of defense submitted in Brazilian labor-inspection proceedings (hereafter, administrative defenses). The main contribution is a transparent and auditable pipeline that combines semantic segmentation, vector retrieval, domain-specific legal embeddings, large language models, similarity-based validation, and human verification to improve traceability and reduce hallucinations. The system was evaluated on real defense documents and expert-provided ground truth using BERTScore, cosine similarity, and DeepEval-based qualitative criteria, with GPT-4o excluded from the DeepEval comparison to avoid judge-model bias. The results show that specialized legal embeddings improve argument identification, with STJ IRIS achieving the highest average cosine similarity among embeddings (0.850). In the full evaluation, the best BERTScore F1 score (0.871) was obtained by GPT-4o; when the evaluation was restricted to validated arguments, Llama 3.3 70B and Llama 4 Scout 17B reached an F1 score of 0.925. These findings indicate that RAG combined with legal embeddings and semantic validation can support reliable legal argument extraction.*

1. Introduction

Large language models (LLMs) are increasingly applied to tasks requiring natural language understanding and generation, reasoning, synthesis, and text-based decision-making [Chen et al. 2024, Lee et al. 2025], with applications in healthcare, finance, and law [Busch et al. 2025, Shool et al. 2025].

In the legal sector, characterized by high document density and highly specialized technical language, LLMs have been used for contract analysis, drafting assistance, and automation of repetitive tasks. However, despite their effectiveness, these models still face a significant challenge: hallucination, that is, the generation of factually incorrect or non-existent content. In this context, where precision and fidelity to the original text

are indispensable, such failures can seriously undermine the reliability of their outputs [Magesh et al. 2025].

The present study aims to evaluate the applicability of the Retrieval-Augmented Generation (RAG) technique to the analysis of administrative defenses filed against infraction notices issued by labor inspectors. This review procedure, regulated by the Ministry of Labor and Employment, reassesses the issuance of an infraction notice based either on the respondent's formal submission or on the administration's duty to review its own acts.

When the respondent submits a defense, the proposed workflow begins by automatically extracting its arguments and validating them against excerpts from the original document. This can significantly accelerate the analytical work and subsequent issuance of opinions, particularly in high-demand contexts.

This article proposes the design and implementation of a RAG-based system that extracts legal arguments from defense documents and validates them through semantic similarity. Unlike generic approaches, the proposed system is tailored to handle the technical vocabulary of Brazilian law and uses specialized legal embeddings and semantic text segmentation. Additionally, the system incorporates mechanisms for traceability and human verification of LLM outputs

To guide the reader, Section 2 reviews related work. Section 3 describes the comparative evaluation of LLMs, system implementation, and assessment procedures. Section 4 discusses the results, and Section 5 presents conclusions and future research directions.

2. Related Work

The combination of artificial intelligence, language models, and legal scholarship has garnered growing attention from researchers, particularly given the potential to automate the reading and analysis of complex texts, such as legal documents, and to assist in their generation. The present work is situated within this interdisciplinary field by proposing the application of the Retrieval-Augmented Generation (RAG) technique to argument extraction from administrative defenses, a topic that remains underexplored in Brazilian scholarship.

International studies have investigated the risks of hallucination when LLMs are applied in the medical and legal domains and have proposed the integration of RAG as a way to mitigate this problem. These studies found that, without an external verification source, the models tend to provide responses that, although fluent and well structured, lack factual support [Rawte et al. 2023].

The use of RAG for legal question answering based on legal documents has also shown promise. Case-based reasoning has been integrated into a RAG pipeline that employs both general-purpose and domain-specific embeddings to retrieve relevant cases and enrich the prompts submitted to large language models, with domain-specific representations improving the quality of the generated answers [Wiratunga et al. 2024]. In a complementary direction, an interpretable retrieve-then-read pipeline for long-form legal question answering has shown that grounding the model in retrieved statutory passages increases the fidelity of the responses, although interpretative errors persist, particularly in identifying the correct legal context [Louis et al. 2024].

In the Brazilian context, studies of the potential and limitations of using LLMs in the legal domain highlight that, despite the sophistication of these models, their use still requires careful human supervision, especially in the drafting of legal opinions [Adriano Fabre et al. 2024]. Objective criteria have also been proposed for assessing the argumentative reliability of AI-generated responses, drawing on real legal documents and semantic similarity analysis [Coelho 2024].

Within the field of automatic verification, different metrics are explored for assessing the quality of responses generated by LLMs, with emphasis on the use of BERTScore and cosine similarity as indicators of coherence and semantic proximity between the generated output and the source text [Juvekar and Purwar 2024]. These metrics are also used in the present study to quantify the completeness and correctness of the responses.

More recently, the performance of legal embeddings on specialized datasets was investigated, and researchers proposed models trained specifically for legal vocabulary [Chang et al. 2024]. Among the embeddings evaluated in this study (Section 4), the legal-domain model STJ IRIS [Melo et al. 2023], built on Legal-BERTimbau, achieved the highest cosine similarity to the expert ground truth, outperforming the general-purpose embeddings MiniLM V6 and Nomic V2 in the identification of technical legal arguments.

Critical analyses also warn against the uncritical use of language models in the production of legal opinions and decisions, particularly because of risks related to bias reproduction and algorithmic opacity [Bruno et al. 2023]. This warning justifies this study's methodological choice to adopt a transparent architecture, with explicit retrieval of the source excerpts that support the extracted arguments.

The present study broadens the range of computational solutions applied to the legal domain by proposing a replicable, auditable, and verifiable system that integrates language models, legal embeddings, and semantic validation to extract arguments from real documents.

3. Methodology

The development of the proposed system for the automated extraction of legal arguments from administrative defenses was based on the Retrieval-Augmented Generation (RAG) architecture and followed an experimental approach structured in four main stages: (i) comparative evaluation of different large language models (LLMs), (ii) implementation of the extraction system, (iii) evaluation of the generated arguments, and (iv) critical analysis of the results.

3.1. Comparative Evaluation of Different Large Language Models (LLMs)

In the first stage, a comparative experiment was conducted across several large language models (LLMs), using a corpus composed of real administrative defenses provided by a subject-matter expert. To ensure methodological consistency, all models were evaluated using the same standardized prompt[1]. The prompt was not subjected to systematic optimization and was not adapted to different types of labor-inspection violations; this controlled choice improves comparability across models but may limit performance in more complex or infraction-specific scenarios. The models included GPT-4o (OpenAI),

[1]https://doi.org/10.5281/zenodo.20796029

Claude 3.7 Sonnet (Anthropic), Gemini 2.5 Pro and Gemini 2.0 Flash (Google), Llama 3 70B (Meta), and DeepSeek R1 with reasoning enabled and disabled.

Modern large language models are already capable of performing argument extraction without the aid of external techniques. However, they may hallucinate and include arguments in their outputs that do not actually exist in the original defense document. To ensure a fair and consistent evaluation, the following standardized procedure was adopted:

- Using a single standardized prompt across all models.
- Ensuring that no model retained conversational context from previous tasks.
- Supplying the defense text in a plain, uniform format to avoid inconsistencies arising from different PDF extraction structures.
- Validating the arguments extracted by each model against the ground-truth arguments provided by a legal expert, using cosine similarity computed with four distinct embeddings.
- Conducting a qualitative evaluation by a domain expert.

3.2. Implementation of the Extraction System

During implementation, we developed a system that integrates natural language processing, semantic analysis, and structured data storage with four core functions: automatic argument extraction, validation of the extracted arguments, traceability with human validation, and maintenance of a structured historical record.

Argument extraction is performed directly on the defense documents in PDF format by large language models using prompts tailored to the Brazilian legal context. The models return the candidate arguments contained in each defense. To ensure that these arguments correspond to the original document rather than to model hallucinations, each candidate is validated by semantic matching: the document is semantically segmented, its segments are encoded using domain-specific legal embeddings (BERT Legal Portuguese), and the cosine similarity between every extracted argument and the document segments is computed. Based on these similarity scores, each argument is automatically classified into one of three confidence levels (Validated, Needs Review, or Not Found), according to user-defined thresholds. In this study, a cosine-similarity threshold of 0.75 was selected empirically to classify an argument as validated. No sensitivity analysis of alternative thresholds was conducted, so the effect of this choice on precision, recall, F1 score, and the number of validated arguments remains a limitation of the evaluation.

Beyond automatic validation, the system is designed to keep a human in the loop and to preserve full traceability. For every match, it records the corresponding similarity score and the specific document excerpts that support the argument, allowing the analyst to inspect the textual evidence directly. The analyst can also confirm or reject each argument through an assisted interface and manually add arguments that were not automatically detected, so that the final set reflects expert judgment rather than the model's output alone.

Finally, every extraction is stored in a relational database as a persistent and auditable record. Each entry retains complete metadata (the model used, the system version, the user, and the case identifier) and can be retrieved through an interface that supports querying and filtering for subsequent analysis. The explicit version control associated with each extraction ensures that results can be reproduced and compared across different versions of the system.

Technical Highlights. The proposed system combines multiple complementary technological components:

- Specialized legal embeddings for contextual understanding;
- Semantic chunking for intelligent text segmentation;
- Cosine similarity validation for argument verification;
- Interactive interface for human-in-the-loop review;
- Structured storage for auditability and analytical processing.

The overall goal is to streamline the review and analysis of legal arguments by automating the identification, validation, and organization of these arguments. This approach reduces manual workload while improving the precision and reproducibility of the identification of argumentative structures within legal documents. The operational architecture of the system is illustrated in Figure 1.

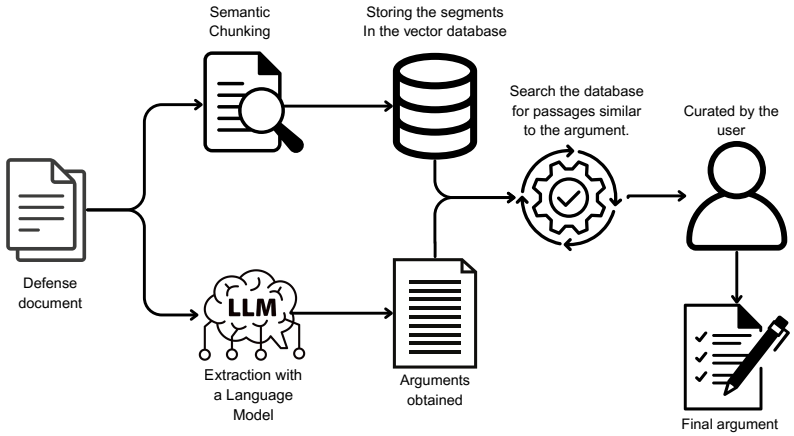


Figure 1. Architecture of the proposed system

3.3. Evaluation of the Generated Arguments

Quantitative evaluation was performed using objective semantic similarity metrics. The methods were BERTScore [Zhang et al. 2020], which measures contextual similarity between generated and reference arguments, and cosine similarity, which measures the similarity between their embedding vectors [Juvekar and Purwar 2024]. To enhance the robustness of the analysis, four distinct embedding models were used: STJ IRIS (specific to the legal domain), MiniLM V6 (a widely used general-purpose embedding), Nomic V2 (an open-source model), and LaBSE (a multilingual model developed by Google).

The qualitative analysis was divided into two complementary components. First, an automated evaluation using the DeepEval library assessed two key criteria: completeness (the presence of all expected arguments) and correctness (the absence of hallucinations or distortions). Second, a legal expert examined aspects such as legal soundness, textual coherence, clarity, and terminological precision. Because only one expert participated, inter-rater agreement could not be measured.

3.4. Critical Analysis of the Results

The arguments generated by the system were assessed using the metrics defined above: BERTScore and cosine similarity for the quantitative dimension, and the automated and

expert evaluations for the qualitative one. For each metric, the analysis reports aggregated mean scores. Measures of dispersion and statistical significance were not computed because the individual results required for those analyses were not available in the consolidated evaluation dataset.

The quantitative scores were complemented by manual inspection of a representative sample of the generated responses. This inspection aimed to capture aspects that automated metrics may miss, such as argumentative coherence, appropriate legal terminology, and completeness with respect to the source defense. This combined quantitative–qualitative analysis supports the assessment of the most suitable chunking and embedding strategies, as well as the limitations discussed in Section 4.

4. Results

The experiments used a corpus of defense documents that a legal expert considered complex and from which language models had been unable to extract the arguments accurately. We first examined how commonly used language models performed at argument extraction and then developed the extraction system to determine whether it outperformed the baseline models.

4.1. Comparative Evaluation of Different Large Language Models

The initial purpose of this stage was to analyze how different embedding models performed when applied to outputs generated by models from different providers. Therefore, this first comparative evaluation included both proprietary and open-source models, enabling a broader analysis of the interaction between model providers, model architectures, and embedding-based similarity metrics.

The arguments obtained from each model were evaluated using cosine similarity computed with four different embedding models. This enabled a broader analysis, as each embedding model was trained on a different text corpus. The embeddings used included STJ IRIS, MiniLM V6, Nomic V2, and LaBSE-Google, thereby encompassing both general-purpose and domain-specific models, such as STJ IRIS, which was trained on legal data.

Table 1 summarizes the aggregated results, facilitating comparison of the models' performance in argument extraction from a diverse corpus of defenses. The data reveal substantial performance differences associated with the models' architectures and the domains of the data used in their training.

The analysis of Table 1 allows for the identification of three relevant patterns for the proposed objective:

1. **Superiority of the specialized embedding:** STJ IRIS, trained on legal data, yielded the highest average cosine similarity scores across the evaluated models (mean of 0.850), supporting the hypothesis that domain-specific embeddings capture argumentative nuances in legal contexts more effectively.
2. **Variation by embedding:** The relative performance of the models changes substantially depending on the embedding used. For example, Claude 3.7 Sonnet performs poorly with Nomic V2 (0.646) but achieves competitive results with STJ IRIS (0.844), suggesting that the choice of embedding directly affects the evaluation.

Table 1. Average cosine similarity between each model's extracted arguments and the expert ground truth, computed with four embeddings (STJ IRIS, MiniLM V6, Nomic V2, and LaBSE-Google). Best result per embedding in bold.

Model	STJ IRIS	MiniLM V6	Nomic V2	LaBSE-Google
GPT-4o	**0.861**	**0.774**	**0.724**	0.821
DeepSeek R1 (reasoning)	0.848	0.715	0.660	0.774
DeepSeek R1 (no reasoning)	0.835	0.721	0.669	0.795
Llama 3 70B	0.846	0.760	0.689	**0.822**
Gemini 2.5 Pro	0.857	0.760	0.721	0.802
Gemini 2.0 Flash	0.859	0.754	0.713	0.793
Claude 3.7 Sonnet	0.844	0.724	0.646	0.786

3. **Competition among proprietary models:** GPT-4o and Gemini 2.5 Pro showed differences of less than 0.05 for three of the four embeddings, indicating that state-of-the-art commercial models achieve similar performance levels, with some variation in specific tasks.

These results demonstrate that model behavior depends not only on architecture but also on the interaction between pretraining and the chosen evaluation metric. As expected, STJ IRIS proved particularly effective for legal applications, while general-purpose, large-scale models, such as Llama 3 70B, can serve as viable alternatives when combined with appropriate techniques.

Although similarity metrics provide quantitative insight, they do not fully capture the quality of model responses. The qualitative evaluation conducted by a legal expert revealed behavioral patterns not evident in the numerical data, focusing on accuracy, completeness, and adherence to prompt instructions.

The most common errors were categorized as follows:

– **Argument hallucination:** The model generates an argument that does not exist in the original text. This occurred in some models, particularly when processing longer texts, as they attempted to "complete" defenses with non-existent arguments.

– **Interpretation vs. extraction:** Some models tended to summarize or reinterpret arguments instead of transcribing them, violating the instructions in the prompt. Although the result was semantically close (hence a good cosine similarity score), it failed to meet the requirement of literal extraction.

– **Argument repetition:** Models frequently repeated arguments that they had already listed earlier.

This qualitative analysis reveals that, for legal use cases where literal accuracy is crucial, similarity scores should be complemented with manual evaluation to ensure that the outputs are not only semantically similar to the source arguments but also faithful to the task specified in the prompt.

4.2. Quantitative Evaluation of Arguments Generated by the System

Unlike the exploratory comparison presented in Section 4.1, the extraction experiments conducted with the proposed system prioritized open-source models because of their

lower operational cost and greater feasibility for repeated use in a practical legal-analysis pipeline. For this reason, the set of models evaluated in the following tables differs from that used in Table 1: the first stage compared models from multiple AI providers to study embedding behavior, whereas the system-level evaluation focused mainly on cost-effective open-source alternatives, while retaining GPT-4o as a proprietary reference baseline.

Using the proposed system, arguments were extracted from defenses considered complex by a legal expert. The resulting set included both correct and incorrect candidate arguments; the latter either fell below the threshold or were not considered genuine arguments. The quantitative metrics were calculated by comparing the arguments extracted by the system with the "ground truth" provided by the expert. The metrics used were:

– **BERTScore:** Precision, recall, and F1 scores computed using the ModernBERT-base embedding model, which allows BERTScore computation with texts longer than 512 tokens.

– **Cosine similarity:** Scores computed using the four embedding models previously mentioned (STJ IRIS, MiniLM V6, Nomic V2, and LaBSE-Google).

The evaluations were conducted using two datasets: one that included both validated and rejected arguments, and another that considered only validated arguments (above the user-defined threshold), thereby comparing extraction with and without the proposed validation strategy.

Table 2. Average BERTScore precision, recall, and F1 scores per model, considering both validated and rejected arguments. Best result per metric in bold.

Metric	DeepSeek R1	Gemma 2	Llama 3.3 70B	Llama 4 Scout 17B	Qwen3 32B	GPT-4o
Precision	0.821	0.820	0.852	0.841	0.811	**0.862**
Recall	0.826	0.831	0.858	0.862	0.807	**0.880**
F1 score	0.824	0.825	0.855	0.851	0.809	**0.871**

Table 3. Average cosine similarity per model and embedding, considering both validated and rejected arguments. Best result per embedding in bold.

Embedding	DeepSeek R1	Gemma 2	Llama 3.3 70B	Llama 4 Scout 17B	Qwen3 32B	GPT-4o
STJ IRIS	0.798	0.850	**0.894**	0.853	0.627	0.838
MiniLM V6	0.710	0.780	**0.808**	0.765	0.497	0.776
Nomic V2	0.703	0.766	0.750	**0.777**	0.591	0.739
Labse - Google	0.756	0.811	**0.812**	0.755	0.748	0.781

Tables 2 and 3 present the evaluation considering the full dataset. All models achieved high BERTScore values for precision, recall, and F1 score, all above 0.81, indicating a strong ability to generate texts that are semantically aligned with the reference arguments. The GPT-4o model achieved the highest F1 score (0.871), showing the best balance between precision and recall. Llama 3.3 70B ranked second in F1 score and achieved the highest cosine similarity score with the STJ IRIS embedding (0.894). Conversely, Qwen3 32B exhibited the lowest performance in this more complex scenario. Since no statistical significance test was conducted, these differences are interpreted descriptively.

Table 4. Average BERTScore precision, recall, and F1 scores per model, considering only validated arguments. Best result per metric in bold.

Metric	DeepSeek R1	Gemma 2	Llama 3.3 70B	Llama 4 Scout 17B	Qwen3 32B	GPT-4o
Precision	0.871	0.872	**0.915**	0.913	0.868	0.911
Recall	0.878	0.888	0.935	**0.937**	0.876	0.929
F1 score	0.875	0.880	**0.925**	**0.925**	0.872	0.920

Table 5. Average cosine similarity per model and embedding, considering only validated arguments. Best result per embedding in bold.

Embedding	DeepSeek R1	Gemma 2	Llama 3.3 70B	Llama 4 Scout 17B	Qwen3 32B	GPT-4o
STJ IRIS	0.819	0.860	**0.893**	0.854	0.790	0.838
MiniLM V6	0.710	0.801	**0.808**	0.765	0.628	0.776
Nomic V2	0.708	0.767	0.754	**0.777**	0.725	0.739
Labse - Google	0.746	0.813	0.812	0.755	**0.840**	0.781

In Tables 4 and 5, the scores remain high and improve overall, as expected when responses below the threshold are excluded. The Llama family models remain in the lead, with Llama 3.3 70B and Llama 4 Scout 17B tied for F1 score (0.925). An interesting result appears for cosine similarity computed with LaBSE: Qwen3 32B ranks first with a score of 0.840. This suggests that, although Qwen3 32B may struggle in more complex full-data scenarios, it performs better when the evaluation is restricted to validated arguments, as reflected in the increases across all similarity metrics.

4.3. Qualitative Evaluation of Arguments Generated by the System

To complement the quantitative metrics based on semantic similarity, an automated qualitative evaluation was conducted with the DeepEval library, using GPT-4o as the judge model. Two fundamental criteria were applied: completeness, which assesses the presence of all relevant arguments expected in the output text; and correctness, which verifies the fidelity of the generated arguments to the source document, penalizing hallucinations—i.e., unsupported or non-existent information in the source text. Because the automated qualitative evaluation uses GPT-4o as the judge model, the GPT-4o extraction model was assessed with the quantitative metrics (BERTScore and cosine similarity) but excluded from the DeepEval-based correctness and completeness comparison. This exclusion avoids self-preference bias because using a model from the same family for both extraction and judging could unfairly favor it.

Table 6. Average correctness and completeness scores per model, considering both validated and rejected arguments. Best result per metric in bold.

Metric	DeepSeek R1	Gemma 2	Llama 3.3 70B	Llama 4 Scout 17B	Qwen3 32B
Correctness	0.636	0.693	0.712	**0.743**	0.608
Completeness	**0.840**	0.764	0.819	0.801	0.781

The results shown in Table 6 indicate that, among the models included in the DeepEval comparison, Llama 4 Scout 17B achieved the highest correctness score when both validated and rejected arguments were considered, whereas DeepSeek R1 achieved the

Table 7. Average correctness and completeness scores per model, considering only validated arguments. Best result per metric in bold.

Metric	DeepSeek R1	Gemma 2	Llama 3.3 70B	Llama 4 Scout 17B	Qwen3 32B
Correctness	0.650	0.677	**0.767**	0.729	0.688
Completeness	0.811	0.776	**0.823**	0.722	0.700

highest completeness score. Table 7 presents model performance after system filtering; in this case, Llama 3.3 70B achieved the highest mean score for both metrics, demonstrating the importance of curation in making publicly accessible models more reliable for legal argument extraction. Because no significance test was conducted, small differences between models should be interpreted descriptively rather than as conclusive evidence of superiority.

These findings underscore the importance of combining quantitative and qualitative evaluation techniques in validating LLM-based systems for legal analysis, particularly in contexts that demand textual rigor and documentary fidelity.

The qualitative analysis conducted by a legal expert revealed significant limitations in the evaluated models that were not detected by the preceding automated metrics. This reinforces the importance of human validation in legally sensitive contexts, such as the automated extraction of arguments in administrative defenses.

Among the main errors observed, the most prominent was argument hallucination—the generation of justifications not present in the original defense text. This type of error severely compromises the legal reliability of the system, as it introduces unsupported elements. Another issue identified was the improper interpretation of arguments: although semantically plausible, some models produced overly simplified or inaccurate summaries, leading to distortions of the original legal content.

Furthermore, violations of the prompt instructions were observed. In several cases, models rewrote arguments, altered enumeration structures, or added explanatory comments, contrary to the explicit instruction to reproduce only the extracted excerpts verbatim.

Despite these limitations, most models performed well when the task was restricted to literal extraction without interpretative interference. This demonstrates that, when guided by well-structured prompts, language models can produce more legally reliable outputs. Thus, the evaluation confirms the relevance of prompt engineering as an essential tool for ensuring control and trustworthiness in text generation for legal applications.

5. Conclusion

This study aimed to explore the application of the Retrieval-Augmented Generation (RAG) technique to the task of automatically extracting legal arguments from administrative defenses, particularly in the context of labor infraction cases. The proposed methodological approach consisted of combining large language models (LLMs) with specialized legal embeddings and semantic validation techniques, enabling not only the generation but also the verification and traceability of the extracted arguments.

Quantitative results showed that domain-specific embeddings, such as STJ IRIS, led to more accurate argument identification than general-purpose embeddings, as reflected

in the highest average cosine similarity scores reported in Table 1. In the exploratory comparison, proprietary models such as GPT-4o and Gemini 2.5 Pro exhibited high similarity between the extracted arguments and the reference arguments. In the system-level evaluation, however, open-source models were prioritized because of their lower operational cost, and models such as Llama 3.3 70B and Llama 4 Scout 17B achieved competitive results after similarity-based validation, as measured by metrics such as BERTScore [Zhang et al. 2020] and cosine similarity [Juvekar and Purwar 2024].

Qualitative evaluation revealed that although LLMs can achieve high accuracy in extraction tasks, they remain susceptible to hallucinations and misinterpretations, especially when exposed to long and complex documents. This finding reinforces the importance of the architecture adopted in this work, which emphasizes explicit similarity-based validation and supports human review with full traceability.

By proposing a practical, transparent, and auditable solution, this study contributes to the advancement of intelligent automation in the legal domain, aligning technical precision with the reliability and security requirements inherent to administrative and judicial practice. Although the use of a standardized prompt reduced interpretation errors, improvements can be achieved through adaptive, infraction-specific prompt engineering, thereby improving the robustness of extraction.

Future research may explore more recent or customized embeddings based on regional case law, Brazilian legal doctrine, or administrative regulations, further optimizing semantic retrieval. Future research should also evaluate RAG-specific metrics such as Contextual Recall, Contextual Precision, Contextual Relevancy, and Faithfulness. Answer Relevancy requires cautious interpretation in this setting because the task is argument extraction rather than question answering. Further experiments should include a sensitivity analysis of the semantic-validation threshold, measures of dispersion, tests of statistical significance, and evaluations by multiple legal experts to quantify inter-rater agreement. A promising direction involves the use of autonomous agents with both short-term and long-term memory, capable of dividing documents into segments, extracting arguments, retrieving related passages, and iteratively reevaluating inconsistencies.

This work may be extended to build a comprehensive RAG-based system that handles multiple documents and links the text of the opinion to applicable legal rules and submitted defenses, thereby supporting the assisted drafting of well-grounded administrative decisions based on precedent and case law.

References

Adriano Fabre, A., Levet Rivera, C. E., Morales Hernández, I. M., and Morales Hernández, J. E. (2024). La transformación de la práctica legal con la inteligencia artificial. *Interconectando Saberes*, 17:133–143.

Bruno, A., Mazzeo, P. L., Chetouani, A., Tliba, M., and Kerkouri, M. A. (2023). Insights into classifying and mitigating llms' hallucinations.

Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H., Kader, R., Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., et al. (2025). Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L., and Wang, W. Y. (2024). A survey on large language models for critical societal domains: Finance, healthcare, and law.

Coelho, A. Z. (2024). A transformação digital e o uso de técnicas inteligência artificial (ia) no sistema de justiça do brasil. *e-Publica*, 11(1).

Juvekar, K. and Purwar, A. (2024). Cos-mix: Cosine similarity and distance fusion for improved information retrieval.

Lee, J., Stevens, N., and Han, S. C. (2025). Large language models in finance (finllms). *Neural Computing and Applications*, 37(30):24853–24867.

Louis, A., van Dijck, G., and Spanakis, G. (2024). Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2025). Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Melo, R., Santos, P. A., and Dias, J. (2023). A semantic search system for the supremo tribunal de justiça. In Moniz, N., Vale, Z., Cascalho, J., Silva, C., and Sebastião, R., editors, *Progress in Artificial Intelligence*, pages 142–154, Cham. Springer Nature Switzerland.

Rawte, V., Sheth, A., and Das, A. (2023). A survey of hallucination in large foundation models.

Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., and Tara, M. (2025). A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., Weerasinghe, R., Liret, A., and Fleisch, B. (2024). Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. In Recio-Garcia, J. A., Orozco-del Castillo, M. G., and Bridge, D., editors, *Case-Based Reasoning Research and Development*, pages 445–460, Cham. Springer Nature Switzerland.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.