REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

USING GENERATIVE AI FOR CLASSIFICATION OF LEGAL DOCUMENTS

UTILIZANDO IA GENERATIVA PARA CLASSIFICAÇÃO DE DOCUMENTOS JURÍDICOS

USO DE IA GENERATIVA PARA LA CLASIFICACIÓN DE DOCUMENTOS LEGALES

Ruan Dias Santana
Bacharel em Ciência da Computação-UFT
Universidade Federal do Tocantins (UFT), Brasil
E-mail: ruan.dias@uft.edu.br

Marcelo da Silva Lisboa
MBA em Data Science–Uninassau
Secretaria de Administração do Tocantins (SECAD), Brasil
E-mail: marcelo.lisboa@secad.to.gov.br

Silvanete Maria da Silva
MBA em Literatura e Linguagem–Cesgranrio
Secretaria de Educação do Tocantins (SEDUC), Brasil
E-mail: silvanete.silva@seduc.to.gov.br

Gabriel Reis Nadler Prata
Graduado em Data Science-Uninassau
Universidade Federal do Tocantins (UFT), Brasil
E-mail: gabriel.prata@uft.edu.br

Marcelo Lisboa Rocha
Doutor em Engenharia Elétrica–UFRJ e Posdoc em Modelagem Computacional-UERJ
PPG em Governança e Transformação Digital - Universidade Federal do Tocantins (UFT), Brasil
E-mail: mlisboa@uft.edu.br

Abstract

The Brazilian judicial system is currently overwhelmed by an enormous backlog of digital lawsuits, making manual case sorting both financially draining and unreliable. This research explores the integration of Generative Artificial Intelligence to streamline the categorization of legal petitions through Large Language Models (LLMs). The study outlines a technical progression divided into three distinct phases. First, a few-shot learning model was tested, resulting in a modest accuracy

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

rate of 56%. Second, the methodology was improved using prompt engineering combined with N-gram analysis and data augmentation strategies to address the issue of skewed datasets. Finally, the research implemented a Retrieval-Augmented Generation (RAG) framework to optimize performance. Using real-world data from the Court of Justice of Tocantins, the experiments demonstrated that the RAG-based system achieved a significant 84% accuracy across 11 complex legal categories. This advanced architecture effectively minimized the occurrence of AI hallucinations and clarified semantic uncertainties often found in legal texts. The findings suggest that this innovative approach provides a reliable and scalable framework for the LegalTech industry, offering a viable path toward modernizing judicial administration. By automating the initial stages of case management, the proposed solution not only enhances operational efficiency but also ensures a higher degree of consistency in the processing of legal documents, ultimately contributing to a more agile and responsive justice system in Brazil and potentially other jurisdictions facing similar digital challenges.
Keywords: Judicial Efficiency; Legal Technology; Artificial Intelligence in Law; Procedural Automation; Digital Justice.

Resumo

O sistema judiciário brasileiro está atualmente sobrecarregado por um imenso acúmulo de processos digitais, tornando a triagem manual de casos um processo dispendioso e pouco confiável. Esta pesquisa explora a integração da Inteligência Artificial Generativa para otimizar a categorização de petições judiciais por meio de Grandes Modelos de Linguagem (LLMs). O estudo descreve uma progressão técnica dividida em três fases distintas. Primeiramente, um modelo de aprendizado com poucos exemplos foi testado, obtendo uma taxa de acerto modesta de 56%. Em seguida, a metodologia foi aprimorada utilizando engenharia de prompts combinada com análise de N-gramas e estratégias de aumento de dados para lidar com o problema de conjuntos de dados enviesados. Finalmente, a pesquisa implementou uma estrutura de Geração Aumentada por Recuperação (RAG) para otimizar o desempenho. Utilizando dados reais do Tribunal de Justiça do Tocantins, os experimentos demonstraram que o sistema baseado em RAG alcançou uma acurácia significativa de 84% em onze categorias jurídicas complexas. Essa arquitetura avançada minimizou efetivamente a ocorrência de alucinações da IA e esclareceu incertezas semânticas frequentemente encontradas em textos jurídicos. Os resultados sugerem que essa abordagem inovadora fornece uma estrutura confiável e escalável para o setor de LegalTech, oferecendo um caminho viável para a modernização da administração judicial. Ao automatizar as etapas iniciais da gestão processual, a solução proposta não só aumenta a eficiência operacional, como também garante um maior grau de consistência no processamento de documentos jurídicos, contribuindo, em última análise, para um sistema de justiça mais ágil e responsivo no Brasil e, potencialmente, em outras jurisdições que enfrentam desafios digitais semelhantes.
Palavras-chave: Eficiência Judicial; Tecnologia Jurídica; Inteligência Artificial no Direito; Automação Processual; Justiça Digital

Resumen

El sistema judicial brasileño se encuentra actualmente saturado por una enorme acumulación de demandas digitales, lo que hace que la clasificación manual de casos sea financieramente agotadora y poco fiable. Esta investigación explora la integración de la Inteligencia Artificial Generativa para optimizar la categorización de peticiones legales mediante Modelos de Lenguaje Largo (LLM). El estudio describe una progresión técnica dividida en tres fases distintas. En primer lugar, se probó un modelo de aprendizaje de pocos disparos, con una modesta tasa de precisión del 56 %. En segundo lugar, se mejoró la metodología mediante ingeniería rápida combinada con análisis de N-gramas y estrategias de aumento de datos para abordar el problema de los conjuntos de datos sesgados. Finalmente, la investigación implementó un marco de Generación Aumentada por Recuperación (RAG) para optimizar el rendimiento. Utilizando datos reales del Tribunal de Justicia de Tocantins, los experimentos demostraron que el sistema basado en RAG alcanzó una precisión significativa del 84 % en 11 categorías legales complejas. Esta

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

arquitectura avanzada minimizó eficazmente la aparición de alucinaciones de IA y aclaró las incertidumbres semánticas que suelen encontrarse en los textos legales. Los hallazgos sugieren que este enfoque innovador proporciona un marco confiable y escalable para la industria LegalTech, ofreciendo una vía viable para modernizar la administración judicial. Al automatizar las etapas iniciales de la gestión de casos, la solución propuesta no solo mejora la eficiencia operativa, sino que también garantiza un mayor grado de consistencia en el procesamiento de documentos legales, contribuyendo así a un sistema de justicia más ágil y con mayor capacidad de respuesta en Brasil y, potencialmente, en otras jurisdicciones que enfrentan desafíos digitales similares.
Palabras clave: Eficiencia Judicial; Tecnología Legal; Inteligencia Artificial en Derecho; Automatización Procesal; Justicia Digital.

1. Introduction

The transition to digital in the legal ecosystem, frequently referred to as "Justice 4.0," has resulted in an exponential increase in the volume of procedural documents, such as petitions, judgments, and rulings. In Brazil, the Electronic Judicial Process system (PJe/E-Proc) has facilitated access to justice, but has created an operational bottleneck: the human inability to process, analyze, and classify the massive flow of unstructured data at the speed demanded by society.

The manual classification of initial petitions — a critical step where the procedural rules and the court's jurisdiction are defined — is a repetitive, costly task subject to human error. Incorrect classification led to erroneous processing, nullities, and significant delays in the administration of justice.

Traditionally, the automation of this task has been approached by classical Natural Language Processing (NLP) techniques, such as Support Vector Machines (SVM) or, more recently, by Transformer-based models such as BERT (DEVLIN, CHANG, *et al.*, 2019). However, these approaches have a critical limitation: the dependence on large sets of labeled datasets. In the legal domain, the creation of these datasets is extremely costly, requiring annotation by specialized jurists, and not by generic labelers (SHUKLA, GUPTA, *et al.*, 2023).

The emergence of Generative Artificial Intelligence and Large-Scale Language Models (LLMs), such as the GPT and Gemini families, offers a paradigm shift. These models, pre-trained on vast corpora of text, possess generalization capabilities that allow them to perform complex tasks with few examples (close-shot learning), drastically reducing the need for specific training

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

data (BROWN, MANN, *et al.*, 2020).

However, the direct application of LLMs in Law faces challenges such as "hallucination" (generation of plausible false information) and the difficulty in dealing with legal intertextuality and nuances of specific theses of higher courts (STF and STJ) (BENTO e TEIVE, 2023).

This article presents the results of an in-depth investigation into the use of the Google Gemini model for the classification of documents at the Court of Justice of Tocantins (TJTO). The work documents the methodological evolution from a static few-shot learning-based approach to a sophisticated Retrieval Augmented Generation (RAG) architecture. The central objective is to demonstrate how the integration of LLMs with vector knowledge bases can solve problems of semantic ambiguity and class imbalance, offering a scalable and highly accurate solution.

2. Theoretical Framework

This section establishes the conceptual pillars that underpin the proposed methodology, addressing the architecture of LLMs, adaptation strategies, and the RAG technique.

2.1. Large-Scale Language Models

LLMs represent the state of the art in NLP. Based on the Transformer architecture, introduced by Vaswani et al. (2017), these models use self-attention mechanisms to weigh the importance of each word in relation to others in a sentence, regardless of the distance between them. This allows capturing long-term dependencies and complex contexts, essential for the interpretation of long legal texts (MAURITZ, 2018).

LLM training occurs in two phases: self-supervised pretraining on massive volumes of data (where the model learns language structure and world knowledge) and fine-tuning for specific tasks.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

In this work, the Gemini family of models (versions 2.0 and 2.5 Flash-Lite) of Google were used. The choice of these models was primarily motivated by economic viability, given the availability of free access to their API, a determining factor for carrying out experiments with limited resources. Secondarily, the choice is justified by their natively multimodal architecture and optimized context window, allowing the efficient processing of extensive requests without the prohibitive costs associated with competing proprietary models, such as GPT-5 (TEAM, ANIL, *et al.*, 2023).

2.2. Fine-Tuning vs. Few-Shot Learning

The adaptation of LLMs to the legal domain follows two distinct paths:

1. Fine-Tuning: This involves retraining the neural network's weights with a domain-specific dataset. While it results in high performance, it requires thousands of annotated examples and high computational power, and presents risks of "catastrophic forgetting" of prior knowledge.
2. Few-Shot Learning (FSL): The model does not undergo weight updates. The task is taught through instructions and a few examples provided in the context of the prompt (in-context learning).

Given the scarcity of labeled data and the need for agility in prototyping, this work initially opted for FSL, later evolving to RAG, which can be interpreted as a dynamic and scalable version of context-based learning.

2.3. Retrieval Augmented Generation (RAG)

One of the main limitations of LLMs is static knowledge (limited to the training cutoff date) and the propensity for hallucinations when confronted with specific technical domains. RAG (Retrieval-Augmented Generation) mitigates these problems by connecting the LLM to a reliable external knowledge base (LEWIS, PEREZ, *et al.*, 2020).

The RAG workflow in this study occurs in three stages:

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

1. Vector Indexing: Reference legal documents are converted into dense numerical vectors (embeddings) that capture their semantic meaning.
2. Retrieval: When a new petition arrives for classification, it is vectorized, and the system searches the database for the k most mathematically similar documents (usually using cosine similarity).
3. Augmented Generation: The LLM receives the new petition along with the retrieved documents as context. The prompt instructs the model to classify the new case based on the provided precedents.

This approach ensures that the model's decision is based on real and up-to-date data, increasing the explainability and reliability of the system.

3. Methodology

The methodology was developed iteratively and incrementally, divided into three experimental phases designed to overcome the generalization limitations encountered at each stage.

3.1. Data Collection and Preparation

The data used were extracted from the E-Proc system of the Court of Justice of Tocantins (TJTO). All documents underwent an anonymization process (removal of names of parties, lawyers, and CPF numbers) to ensure compliance with the LGPD (Brazilian General Data Protection Law). The "ground truth" classification was performed manually by magistrates and legal advisors.

Three versions of datasets were built throughout the project:

- Dataset Base (dataset-temas-novo.csv): Version used as a baseline, composed of 710 documents distributed in 5 classes: THEME 864, THEME 986, THEME 1118, THEME 1177 and NONE. This set presented a severe imbalance, with the majority classes containing approximately 300 examples and the minority classes (such as THEME 1118) only 21 instances.
- Revised Dataset (arqtemas-ANTIGO01.csv): Improved version with text

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

cleanup and semantic normalization. For the Phase 2 experiments, this dataset was subjected to Data Augmentation techniques (back translation and synonym substitution), expanding all classes to 300 examples each, resulting in a balanced corpus of 1,500 instances.

- Expanded Dataset (arqtemas-NOVO01.csv): Incorporation of unprecedented themes of general repercussion and repetitive appeals from the STF and STJ (THEMES 793, 1184, 1199, 566, 1132 and 796). In the final stage (RAG), the sets were unified, totaling 11 distinct classes to validate the robustness of the model in a scenario of greater legal complexity.

3.2. Definition of Thematic Classes

Each of the thematic classes considered in this work corresponds to a specific case law consolidated by the Brazilian superior courts, more precisely:

- Supreme Federal Court (STF): General Repercussion issues, which establish binding understandings on constitutional matters.
- Superior Court of Justice (STJ): themes of Repetitive Appeals, which standardize the interpretation of infra-constitutional legislation.

These themes therefore function as normative legal categories, into which each text of a legal request may or may not fit according to its content and justification. This characteristic reinforces the interpretative and contextual nature of the classification task: it is not just about identifying isolated words, but about understanding the semantic relationship between the legal request and the corresponding legal argument.

The **NONE** class represents a critical residual grouping, containing requests that do not fit into any of the previously established theses (potential false positives).

3.3. Prompt Engineering and Data Augmentation

In Phase 2, to combat the imbalance, a Hybrid Data Augmentation pipeline was applied:

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

1. Backtranslation: Translation of legal texts into English and re-translation into Portuguese using translation APIs, generating natural syntactic variations while maintaining semantics.
2. Synonym Replacement: Swapping technical terms for equivalents (e.g. "vehicle" for "car", "claimant" by "author"), reduce vocabulary variability by using simpler terms.

Additionally, the prompts were enriched with N-grams. Word sequences (bigrams) were extracted and trigrams) most frequent of each class and inserted in the prompt as explicit "cues" for the model to focus on determining terms.

3.4. RAG Architecture

The definitive solution (Phase 3) abandoned synthetic augmentation in favor of RAG. The following settings were used:

- Embeddings: Google's text-embedding-004 model was used to vectorize 80% of the unified dataset. The corresponding embedding has 768 dimensions.
- Vector Database: The vectors were indexed in the Pinecone, optimized for high-dimensional search.
- Inference Process: For each test document (20% remaining), the system retrieved the k = 5 most similar documents from the training base. The Final Prompt instructed Gemini to classify the target document considering only the evidence present in the retrieved documents.

4. Results and Discussions

The experimental evaluation was conducted sequentially and incremental. This approach allowed us to isolate the contribution of each technique — from simple prompting to retrieval of context — for the effectiveness of legal classification. Next, the quantitative performance and qualitative implications of each experimental phase are discussed.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

4.1. Phase 1: Static Few-Shot Limitations

In the first phase, the baseline was established using the GPT-4o-mini model with static prompts (Few-Shot), without access to external knowledge base. The objective was verifying the intrinsic ability of the model to generalize legal standards only with pre-trained knowledge.

As shown in Table 1, performance was unsatisfactory, with an overall accuracy of only 56%. The model exhibited a strong bias towards the majority class (THEMA 864) and for the residual class (NONE), ignoring the specificities of complex tax issues.

Table 1. Detailed Classification Report – Phase 1 (Baseline)

Class	Precision	Recall	F1-Score	Support
NONE	0.68	0.48	0.57	270
THEME 1118	0.03	0.10	0.04	10
THEME 1177	0.27	0.29	0.28	14
THEME 864	0.56	0.74	0.64	273
THEME 986	1.00	0.03	0.05	37
Accuracy			0.56	604
Macro Average	0.57	0.34	0.33	604
Weighted Average	0.63	0.56	0.56	604

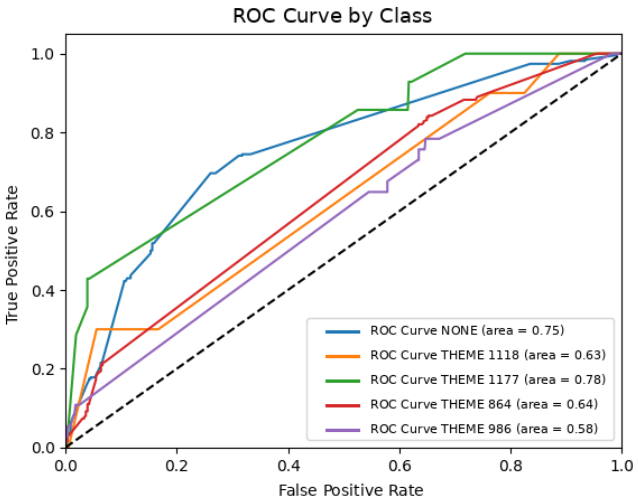
Critical failure occurs in minority classes. THE THEME 986, for example, obtained a Recall of just 0.03, indicating that the model failed to identify 97% of the petitions relating to this theme.

Figure 1 visually corroborates this scenario. The ROC Curve presents a reduced area under the curve (AUC) for unbalanced classes, approaching the diagonal line (randomness). The low convexity of the curves for minority classes indicates weak separation capacity between themes. This highlights that, without context, LLM tends to "hallucinate" or resort to generic statistical probabilities, which is unacceptable in high-precision legal applications.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

Figure 1. ROC curve of Phase 1.



4.2. Phase 2: Impact of Data Augmentation

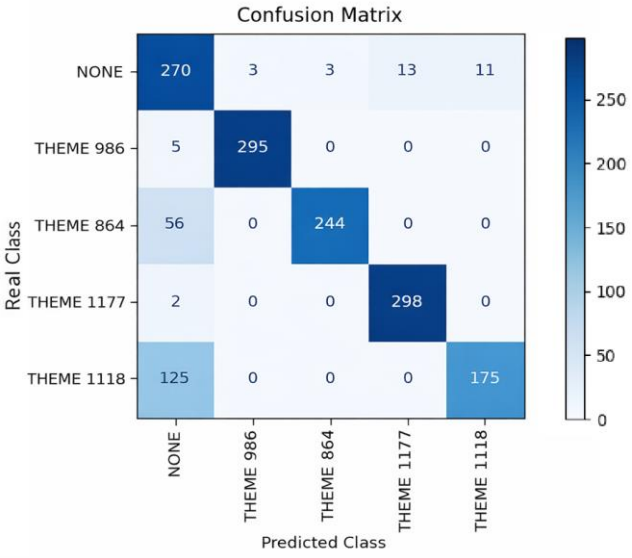
In the second phase, we sought to mitigate the imbalance through Data Augmentation and refinement of prompts with N-grams. This strategy increased accuracy to 85% in Revised Dataset.

The main contribution of this phase was the reduction of false positives in the "NONE" class. The Confusion Matrix in Figure 2 reveals a more defined main diagonal for the original themes. A reduction in dispersion is observed in the NONE class.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

Figure 2. Confusion Matrix - Phase 2 (Revised Dataset).

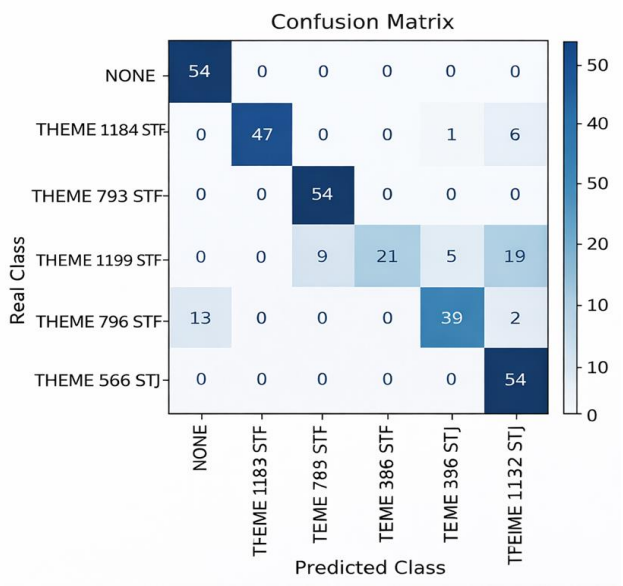


However, expanding to the New Themes Dataset (without the residual class) revealed the limitation of this approach based solely on N-grams: semantic ambiguity. The N-gram approach shows limitations in distinguishing semantically related themes. Figure 3 illustrates how themes that share similar vocabulary (e.g., THEME 796 and THEME 1199) generated confusions that did not exist in the previous scenario.

REMUNOM
ISSN 2178-6925

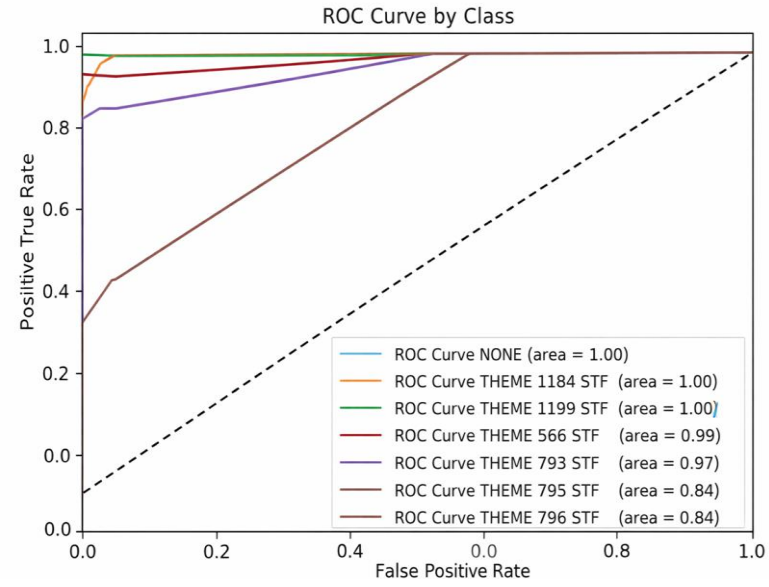
Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

Figure 3. Confusion Matrix - Dataset of New Themes.



The ROC curve for this scenario, as shown in Figure 4, confirms that, although overall accuracy remains high, the ability to discriminate drops for specific classes (such as TEMA 796), indicating the need for a more robust contextual approach (RAG).

Figure 4. ROC Curve - Dataset of New Themes.



REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

4.3. Phase 3: Consolidation with RAG

The implementation of the RAG (Retrieval-Augmented Generation) architecture represented the definitive evolution of the system. By anchoring text generation to excerpts retrieved from real case law, it was possible to perform a stress test with 11 simultaneous classes without the need for synthetic balancing.

Table 2 presents the final consolidated results. Overall accuracy stabilized at 84%. Although numerically similar to the previous phase, this result is qualitatively superior, as it was obtained in a scenario of doubled complexity (11 classes vs. 5 classes).

The performance stands out, with an improved F1-Score compared to the previous Phases 1 and 2, in complex themes such as THEME 864, THEME 1118, THEME 796, and THEME 566. This confirms the hypothesis that information retrieval acts as an "anchoring" mechanism, transforming the creative hallucination task into a semantic verification task.

Table 2. Final Classification Report with RAG (11 Classes)

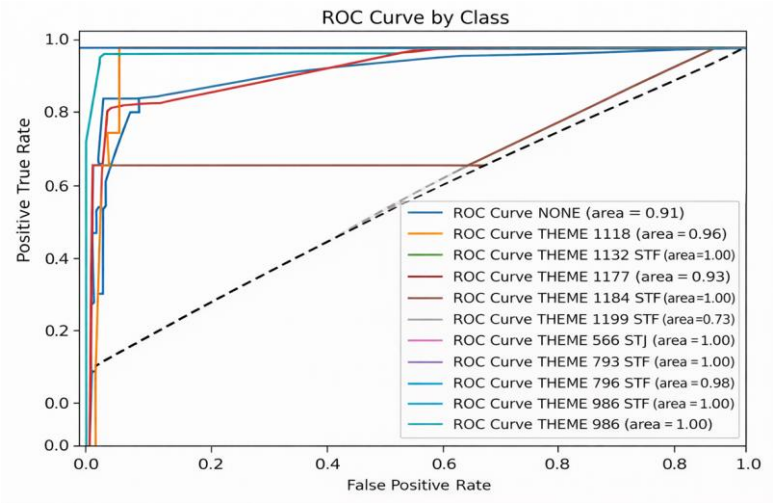
Class	Precision	Recall	F1-Score	Support
NONE	0.90	0.63	0.75	60
THEME 1118	0.21	0.75	0.33	4
THEME 1132 STJ	1.00	1.00	1.00	2
THEME 1177	0.44	0.67	0.53	6
THEME 1184 STF	1.00	1.00	1.00	8
THEME 1199 STF	1.00	0.67	0.80	3
THEME 566 STJ	1.00	1.00	1.00	2
THEME 793 STF	0.92	1.00	0.96	11
THEME 796 STF	1.00	1.00	1.00	1
THEME 864	0.91	0.98	0.94	60
THEME 986	1.00	1.00	1.00	13
Accuracy			**0.84**	170
Macro Average	0.85	0.88	0.85	170
Weighted Average	0.88	0.84	0.85	170

Finally, Figure 5 shows the final ROC curves. The approximation of the curves to the upper left corner for the vast majority of classes validates the robustness of the RAG classifier, making it suitable to assist in the screening of processes on a large scale.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

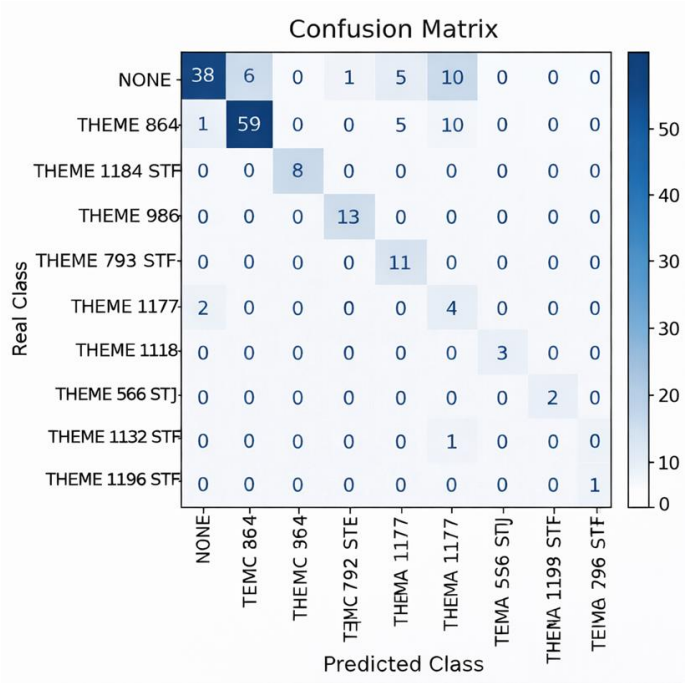
Figure 5. Final ROC Curve (RAG).



4.4. Analysis and Limitations

Despite the overall success, the qualitative analysis of the final Confusion Matrix, as in Figure 6, indicates that residual errors persist in the NONE class. The model still incorrectly classifies some generic petitions as belonging to specific themes that have overlapping vocabulary, notably THEME 1177.

Figure 6. Confusion Matrix - Final RAG.



REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

This occurs because, semantically, the vectors of a "Vehicle Property Tax Fine Enforcement" petition (class NONE) and a "Declaratory Action of Non-Existence of Vehicle Property Tax Debt by Sale" (THEME 1118) are very close in latent space. The retrieval mechanism brings both types of documents, and the LLM, when in doubt, tends to opt for the specific class. This suggests that, for future applications, semantic search alone is not enough; a re-ranking step or an additional logical check after retrieval may be necessary.

5. Conclusions

This work demonstrated the feasibility and effectiveness of using Large-Scale Language Models, specifically Gemini, integrated with a Retrieval Augmented Generation (RAG) architecture, for the automatic classification of judicial documents.

The methodological evolution, starting from 56% accuracy with simple few-shot to 84% with RAG, validates the hypothesis that dynamic context is the key to high-precision legal automation. Data Augmentation and N-grams techniques served as important intermediate steps, but it was the ability of RAG to consult a knowledge base in real time that ensured the final robustness of the classifier in the face of the complexity and intertextuality of Brazilian Law.

In addition to the accuracy metric, the RAG architecture offers a crucial advantage for the public sector: explainability.

Unlike "black-box" models, it is possible to audit exactly which precedents were used by the system to support each classification, increasing confidence in the technology.

As future work, it is suggested to explore embedding models specifically trained on Brazilian legal corpora (such as LegalBERT-PT) to refine the retrieval stage, and to validate the system in a production environment to measure gains in procedural efficiency.

REMUNOM
ISSN 2178-6925

Received: 13/01/2026 - **Accepted:** 20/02/2026
Vol: 02.03
DOI: 10.66104/4zsfgz56
Pages: 1-16

References

BENTO, F. M.; TEIVE, R. C. G. Classificação de documentos jurídicos utilizando a arquitetura transformer: uma análise comparativa com algoritmos tradicionais de Machine Learning e ChatGPT. **Brazilian Journal of Development**, v. 9, p. 20208–20224, 2023.

BROWN, Tom et al. **Language models are few-shot learners**. Advances in neural information processing systems, v. 33, p. 1877-1901, 2020.

DEVLIN, J. et al. **BERT:** Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, V. 1. 2019. p. 4171–4186.

LEWIS, Patrick et al. **Retrieval-augmented generation for knowledge-intensive nlp tasks**. Advances in neural information processing systems, v. 33, 2020.

MAURITZ, Bc Jiří. **Automatic classification of legal documents**. 2018. Tese de Doutorado. Master's thesis, Masarykova univerzita. Available in <https://is.muni.cz/th/kt3kh/thesis_Archive.pdf>

SHUKLA, Bharti et al. **Challenges and issues in legal documents classification**. In: AIP Conference Proceedings. AIP Publishing LLC, 2023.

TEAM, G. et al. Gemini: a family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2023.

VASWANI, A. et al. **Attention is all you need**. Advances in neural information processing systems 30. 2017. p. 5998–6008.