Structuring Information from Initial Petitions Using LLMs: A Study in Brazilian Courts of Justice

Rhedson Esashika[1],
Carlos M. S. Figueiredo[1], **Tiago de Melo**[1]

[1]PPGEEL - Postgraduate Program in Electrical Engineering.
State University of Amazonas (UEA), Manaus, AM, Brazil

{rffe.mee25, cfigueiredo,tmelo}@uea.edu.br

Abstract. *This study proposes the use of Large Language Models to extract structured information from Brazilian legal petitions. We guided four models (from the Gemini and Gemma3 families) to generate structured JSON outputs, providing a comparative performance benchmark for this novel task. The main contribution is a validated workflow, where results show Gemini models significantly outperform Gemma in capturing complex semantic data. This work establishes a robust evaluation and an important methodological baseline for a critical task in Brazilian legal Natural Language Processing (NLP).*

1. Introduction

In the contemporary legal domain, the increasing demand for procedural celerity drives the search for technological solutions capable of optimizing judicial case processing. Advanced Artificial Intelligence (AI) techniques, particularly Large Language Models (LLMs), are emerging as transformative tools for analyzing and processing legal documents [Widyassari et al. 2022]. The task of text summarization, a key technique in Natural Language Processing (NLP), is essential for automatically extracting relevant information from large volumes of textual data while preserving core elements such as facts, arguments, and requests [Supriyono et al. 2024]. In the legal field, the ability to accurately and structurally synthesize petitions represents a strategic advancement for optimizing analytical activities and strengthening access to justice.

Recent years have seen significant advancements in LLMs, driven by large-scale textual data and evolving neural network architectures [Jayatilleke et al. 2024]. These models have markedly improved their capabilities in semantic interpretation, information extraction, and summarization, enabling them to address the complexity of legal documents with greater precision [Hussain and Thomas 2024]. However, applying these technologies to the Brazilian legal context introduces unique challenges, stemming from the intricacies of its judicial system and the specific linguistic characteristics of Portuguese. This scenario needs a focused evaluation of how effectively modern LLMs can be adapted to this national and domain-specific environment.

Aligned with these developments, this study contributes by evaluating the performance of recent LLMs—including Gemini 2.5 Pro, Gemini 2.0 Flash, and Gemma 3—in extracting structured information from Brazilian initial petitions. While the underlying prompt engineering methods are well-established, their application to the unique linguistic and structural complexities of these documents remains an underexplored area. Therefore, this study's primary contribution is not methodological innovation but the validation

and adaptation of these techniques to bridge a critical gap. We propose and validate a structured extraction workflow tailored for the Brazilian legal domain, providing the first comparative analysis of state-of-the-art LLMs on this novel task. Specifically, we assess the models' capacity to identify seven key fields: Plaintiff, Defendant, Key Issues, Facts, Legal Theory, Claims, and Issues for Decision.

The remainder of this paper is structured as follows: Section 2 presents the related work, detailing previous studies on legal information extraction and the application of LLMs in the legal domain. Section 3 describes the materials and methods employed, including the models, datasets, and prompt engineering strategies applied. Section 4 presents and discusses the experimental results, highlighting the performance and limitations of the proposed approach. Finally, Section 5 concludes the study and outlines directions for future research.

2. Related Work

The application of LLMs to extract and structure legal documents has gained attention in recent years. Nevertheless, challenges like textual variability, terminological ambiguity, and a lack of specialized annotations continue to pose significant barriers to the automation of legal information processing automation. Recent studies have investigated extracting legal attributes and mapping unstructured legal texts into structured formats. This section reviews key contributions in these research areas.

The extraction of structured legal attributes has been widely investigated. Adhikary et al. [Adhikary et al. 2024] proposed a weakly-supervised approach based on few-shot prompt labeling to extract thematic attributes from criminal petitions, reducing the need for large annotated datasets. Zin et al. [Zin et al. 2024] focused on extracting normative statements from German judicial decisions using advanced prompting strategies like Chain-of-Instructions (CoI), showing notable improvements even in zero-shot scenarios. Furthermore, Breton et al. [Breton et al. 2025] examined the extraction of legal terms from traffic legislation, utilizing models such as GPT-4 and Mixtral to tackle multi-label classifications and entity overlap. Wan et al. [Wan et al. 2024] introduced an Adapt-Retrieve-Revise framework to improve domain adaptation of LLMs' domain adaptation for legal tasks, demonstrating that retrieval-augmented strategies improve the extraction of structured information from complex judicial documents. These contributions highlight the need for precise extraction techniques to address the complexities of legal language.

Paticularly, mapping unstructured legal texts to structured representations has gained significant interest. Chalkidis et al. [Chalkidis et al. 2021] introduced LexGLUE, a benchmark dataset for evaluating legal NLP models in classification, extraction, and summarization. Breton et al. [Breton et al. 2025] structured extracted legal terms into standardized JSON formats, aiding integration into legal information systems. Ma et al. [Ma et al. 2024] proposed the LLMParser system to convert semi-structured logs into adaptable templates for legal documents. Ngo et al. [Ngo et al. 2023] presented an integrated ontology to represent legal knowledge from legal documents. Their model integrates the ontology of relational knowledge and a graph of key phrases and entities to express the semantics of legal content. These studies highlight the importance of effective structuring methodologies for enhancing automation and analysis of judicial document.

These studies underline the significance of robust structuring methodologies for improving the automation and analysis of judicial processes.

While much of the research focuses on English or other widely used languages, the specific challenges associated with Brazilian Portuguese legal texts are gaining attention. Sakiyama et al. [Sakiyama et al. 2023] examined text decoding strategies for automatically generating legal keyphrases in Brazilian Portuguese, focusing on morphological richness and specialized legal terminology. Silveira et al. [Souza et al. 2023] introduced the LegalBERT-pt model, adapting transformer architectures for the Brazilian legal domain and showcasing the advantages of domain-specific pretraining. Moreover, Aquino et al. [de Aquino et al. 2024] proposed using Retrieval-Augmented Generation (RAG) techniques to extract structured information from Brazilian legal documents on public procurement fraud, demonstrating the effectiveness of hybrid approaches.

While previous studies have developed domain-specific models and case studies for several legal tasks, our work focuses on a practical and specific application: the structured extraction of seven key fields from Brazilian initial petitions. Our method applies a prompt strategy using general-purpose LLMs. These specific key elements are very align to real Tribunal workflows which provides to the judiciary system a benchmark showing the potential and limitations of the approach in terms of efficiency and costs.

3. Methodology

The proposed methodology consists of employing LLMs for the extraction of fields based on legal and informational criteria relevant to the judicial system. The overall workflow of the proposed methodology is illustrated in Figure 1. The pipeline comprises four main stages: (i) **Case Distribution**, which refers to the reception and classification of the initial petition; (ii) **Preprocessing**, where the raw textual content is prepared and normalized; As part of this stage, an analysis of the document lengths was conducted. It was confirmed that all petitions in the dataset were sufficiently short to fit entirely within the context windows of all evaluated models. Therefore, no truncation or text selection methods were necessary.(iii) **LLM + Prompt Engineering**, where the petition is processed through a LLM guided by a persona-based prompt; and (iv) **Structured Output**, which delivers a JSON object containing the extracted legal fields.

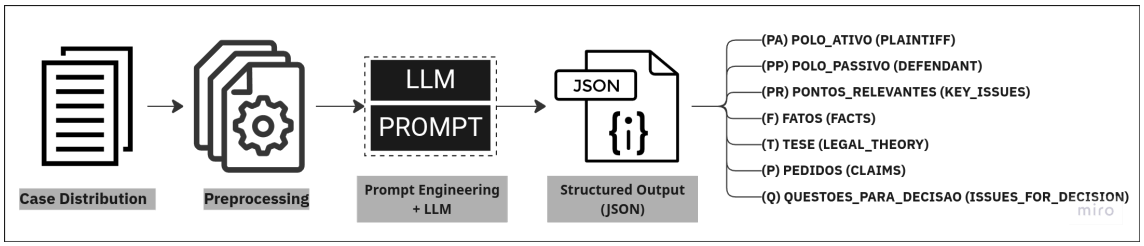

Figure 1. Information extraction pipeline using LLMs.

The following subsections describe the dataset, the models and prompts employed, as well as the experimental protocol adopted for the evaluation of results.

3.1. Dataset

For the intended experiment, a sample of initial petitions was collected from the procedural archive of the Court of Justice of the State of Amazonas, restricted to actions

filed during the month of January 2025. The documents cover two distinct jurisdictional categories—Special Civil Courts and Ordinary Civil Courts—with the aim of ensuring thematic diversity and representativeness of the different levels of legal formalism characteristic of these segments. The resulting dataset consists of 100 initial petitions, of which 50 pertain to the Special Civil Courts and 50 to the Ordinary Civil Courts. The documents have an average length of approximately 2,576 words, reaching a maximum of 6,070 words. We acknowledge that this sample size is modest; however, it is sufficient for the exploratory nature of this study, which aims to provide a comparative performance analysis of recent LLMs on this specific legal task. The dataset constructed for this study, which will be made publicly available, serves as an initial benchmark for future research in this area.

In order to ensure compliance with the regulatory standards and best practices currently established within the Brazilian Judiciary, the present study was conceived in accordance with the principles, guidelines, and restrictions established by Resolution CNJ No. 615, of March 11, 2025 [Conselho Nacional de Justiça 2025]. Besides, in compliance with the provisions of the General Data Protection Law (Law No. 13.709/2018) [Presidência da República 2018], pursuant to Article 7 of the aforementioned Resolution, no information protected by judicial confidentiality was submitted to the models. The dataset constructed for this study will be made publicly available through an online repository. The corresponding URL will be provided upon acceptance of the article for publication. Additionally, the dataset comprises petitions with varied topics and textual structures, encompassing both standardized claims from Special Civil Courts and more elaborate narratives from Ordinary Civil Courts, ensuring coverage across different levels of legal formalism.

3.2. Large Language Models (LLMs) Employed

In this study, four LLMs with distinct architectures and varying levels of computational capacity were employed, as detailed in Table 1, with the objective of evaluating their performance in the task of extracting structured information from initial petitions. For local models we show their size to infer the computational infrastructure needed, while we show the costs of API models[1] to use cloud resources.

Table 1. Applied LLMs with Respective Modes of Access and Costs.

Model	Parameters	Access	Input Tokens 1 mi — USD	Output Tokens 1 mi — USD
Gemini 2.5 Pro	Not disclosed	API	1,25	10,00
Gemini 2.0 Flash	Not disclosed	API	0,10	0,60
Gemma 3 27B IT	27 Billion	Local	free	free
Gemma 3 1B IT	1 Billion	Local	free	free

The Gemini 2.5 Pro (gemini-2.5-pro-exp-03-25) is a state-of-the-art multimodal model developed by Google DeepMind [Google DeepMind 2025], endowed with advanced capabilities in contextual reasoning and semantic understanding across specialized domains. In this study, the experimental version made available in March 2025 through the Vertex AI platform was employed. Complementarily, the Gemini 2.0 Flash,

[1] https://ai.google.dev/gemini-api/docs/pricing?hl=en

an optimized variant designed for lightweight reading tasks, with an emphasis on rapid generation and efficient performance in low-latency scenarios, was also utilized.

Additionally, two models from the Gemma 3 Instruction-Tuned family were employed: gemma-3-1b-it, with approximately 1 billion parameters, and gemma-3-27b-it, with 27 billion parameters [Team et al. 2025]. Both models are open-source and adapted for instruction-driven tasks, with the former suited for environments with constrained computational resources and the latter intended for assessing the impact of scale on the quality of generated responses.

The Gemini variants were accessed via the Vertex AI API, while the Gemma models were executed locally using the LM Studio environment [LM Studio 2025] on a 16-inch MacBook Pro with an Apple M4 Pro chip, featuring an integrated X-core GPU, and 24 GB of unified memory. To ensure consistent and reproducible results, specific generation parameters were maintained across all experiments. The temperature was set to 0.1 and top-p was set to 0.5 for all models. None of the models underwent additional fine-tuning; all interactions were conducted exclusively through the technique of prompt engineering in a few-shot configuration, employing the system prompt described in Section 3.3.

Other LLM families, such as GPT (OpenAI), Claude (Anthropic), and LLaMA (Meta), were not included in this study. The focus was restricted to models from a single ecosystem (Gemini and Gemma) to enable a controlled comparison across architectures and computational configurations. This choice reflects a defined scoping strategy rather than an evaluation of model quality, and it ensures methodological consistency under comparable cost and infrastructure conditions.

3.3. Output Structure and Prompt Engineering

The language models were instructed to generate their outputs in a structured JSON format, developed with the purpose of organizing the main legal elements present in initial petitions. This structure was designed with the potential to facilitate subsequent integration with case management systems and to enable the automation of judicial triage, analysis, and classification stages.

The desired output must contain seven main fields, defined based on legal and informational criteria relevant to the activities of judges, court clerks, and artificial intelligence systems. These fields, in Brazilian portuguese, are described below:

- `POLO_ATIVO`: identification of the plaintiff parties.
- `POLO_PASSIVO`: identification of the defendant parties.
- `PONTOS_RELEVANTES`: enumeration of the key issues (central legal points).
- `FATOS`: chronological summarization of the facts without using lists.
- `TESE`: main legal thesis supporting the judicial request.
- `PEDIDOS`: formulated claims.
- `QUESTOES_PARA_DECISAO`: legal issues that require judicial decision.

In order to guide the language models toward a legally coherent, accurate, and structured extraction of the information contained in the initial petitions, a prompt engineering approach based on a specialized persona and explicit instructions was employed, as illustrated in Figure 2. The prompt was formulated to simulate the legal reasoning of

an experienced judge, incorporating semantic, conceptual, and structural elements appropriate to the domain of Civil Procedural Law.

The formulation establishes a specialized interpretative context in which the model assumes the role of a judge with extensive experience (20 years). The inclusion of the instruction "Think step by step" (*"Pense passo a passo"*) was employed as a chain-of-thought (CoT) induction technique [Wei et al. 2022], aiming to stimulate structured reasoning and the logical decomposition of the extraction task. This strategy has proven effective in generating more cohesive and semantically organized responses, particularly in fields that require contextual inference, such as the chronological reconstruction of legal facts.

An illustrative example was included within the body of the prompt, characterizing a *few-shot learning* approach. This strategy aimed to provide the models with a reference instance regarding the expected structure and content of the responses, promoting greater adherence to the desired legal and semantic format.

PERSONA
- Você é um ****MAGISTRADO****, com mais de 20 anos de experiência, Especialista em SUMARIZAR ****PETIÇÕES INICIAIS**.**

TAREFA
- Extraia em **formato JSON** siga estritamente o a estrutura <estrutura_json>:
- Utilize como exeplo EXEMPLO_JSON
- Pense passo a passo.

ESTRUTURA_JSON
<estrutura_json>
{
"POLO_ATIVO": `caso tenha mais de um, separe por ponto e vírgula ";"`,
"POLO_PASSIVO": `caso tenha mais de um, separe por ponto e vírgula ";"`,
"PONTOS_RELEVANTES": `separe os pontos jurídicos relevantes`,
"FATOS": `sumarize os fatos de forma cronológica, não utilize listas`,
"TESE": `informe a tese jurídica, conceito: fundamento legal central que sustenta o pedido feito ao juiz, resumindo o direito alegado pelo autor na ação judicial. `,
"PEDIDOS": `informe os pedidos da petição inicial, separe por ponto e vírgula ";"`,
"QUESTOES_PARA_DECISAO": `informe as questões jurídicas a serem decididas, conceito: toda questão posta no processo que exige a emissão de um juízo decisório pelo juiz`
}
</estrutura_json>

EXEMPLO_JSON
<exemplo_json>
{
"POLO_ATIVO": "[NOME DA PARTE AUTORA]",
"POLO_PASSIVO": "[NOME DA EMPRESA RÉ]",
"PONTOS_RELEVANTES": "Negativação indevida; Dano moral; Ausência de relação jurídica entre as partes; Pedido de inversão do ônus da prova; Justiça gratuita; Juízo 100% digital; Desinteresse na audiência de conciliação.",
"FATOS": "Em [DATA ESPECÍFICA], o autor teve seu nome negativado pelo réu em decorrência de uma dívida de [VALOR DA DÍVIDA], cuja origem o autor desconhece. O autor afirma não ter contratado nenhum serviço ou produto junto ao réu e que não foi notificado previamente sobre a negativação. Tentou resolver a situação administrativamente, solicitando documentos comprobatórios da dívida, mas não obteve resposta. Diante da negativação, o autor teve seu crédito cerceado. Por fim, solicitou a retirada do nome do cadastro de inadimplentes, sem sucesso.",
"TESE": "O autor alega que foi negativado indevidamente pelo réu, pois não reconhece a dívida e nunca firmou contrato com a empresa. Sustenta que a inscrição indevida em cadastros de proteção ao crédito configura dano moral in re ipsa, sendo o réu responsável objetivamente pela reparação. Aduz ainda a hipossuficiência do consumidor e a dificuldade de comprovar a inexistência da dívida, requerendo a inversão do ônus da prova.",
"PEDIDOS": "Concessão de medida liminar para retirada do nome do cadastro de inadimplentes; Juízo 100% digital; Apresentação do contrato e documentos da dívida pelo réu; Retirada definitiva do nome do cadastro de inadimplentes; Proibição de novas inscrições referentes à mesma dívida; Fixação de multa diária; Dispensa de audiência de conciliação; Justiça gratuita; Citação do réu; Declaração de inexistência de relação jurídica; Indenização por danos morais no valor de [VALOR DA INDENIZAÇÃO]; Inversão do ônus da prova; Correção monetária e juros; Condenação do réu em custas e honorários advocatícios.",
"QUESTOES_PARA_DECISAO": "Existência ou não de relação jurídica entre as partes; Validade da dívida cobrada; Cabimento da inversão do ônus da prova; Configuração de dano moral e seu valor; Concessão ou não da tutela antecipada."
}
</exemplo_json>

Figure 2. Prompt used to obtain the result.

3.4. Evaluation Procedures

In order to perform a comparative evaluation of models, each petition from the collected dataset was processed by the four selected language models using the structured prompts described in Figure 2, totaling 400 distinct executions for the generation of structured outputs.

Accordingly, the full set of 400 generated responses was divided among a team of three judicial analysts with legal training and experience, who volunteered from the Court of Justice. To distribute the workload, each response was assigned to and evaluated by a single analyst from the team. The choice of human evaluation stems from the interpretative nature of the task, in which semantic, legal, and pragmatic aspects are not fully captured by traditional automatic metrics. Thus, an approach centered on the contextual validity of the extracted information was adopted, aiming to reflect the real-world requirements for the practical use of outputs within the justice system. Each response was evaluated by one among three legal analysts, ensuring that the full set was reviewed by multiple qualified professionals.

To better guide the evaluation conducted by the analysts, a scoring protocol was developed, consisting of an ordinal scale with four levels, as detailed in Table 2. Each generated instance was individually examined and classified according to this four-level ordinal scale, ranging from *Irrelevant or Potentially Harmful* (level 1) to *Highly Relevant and Ready for Implementation* (level 4). The objective of this protocol is to measure the legal relevance and the degree of operationalization of the extracted information according to the opinions of judicial system professionals, thereby allowing for the estimation of the practical feasibility of the proposed method for real-world application in the legal context.

Table 2. Evaluation Protocol for the Relevance of Extracted Information.

Level	Classification	Description
1	Irrelevant or Potentially Harmful	Incorrect or out-of-context content with no relation to core procedural elements; may mislead legal professionals or harm automation.
2	Low Relevance or Significant Ambiguity	Related to context, but imprecise, fragmented, or ambiguous; limits practical utility.
3	Relevant and Operationalizable	Contributes significantly to understanding or automation; has practical value.
4	Highly Relevant and Ready for Implementation	Accurate, complete, and directly applicable; high strategic value for case management and AI automation.

For quantitative analysis purposes, the percentage distributions of the classifications assigned to the outputs of each model, evaluated by JSON field, were calculated. Additionally, a simple arithmetic mean was computed for each model, using the ordinal scale from 1 to 4 as numerical representations of the qualitative levels. This aggregated metric enabled a global performance comparison among the different LLMs, allowing for an estimation of the legal relevance and operational feasibility of the extracted information.

4. Results and Discussion

4.1. Presentation of Results

The results obtained were organized into a comparative matrix, as presented in Table 3. Each value corresponds to the average of the scores assigned for each semantic field,

according to the ordinal scale from 1 to 4 described in Table 2, for each evaluated model. Furthermore, the evaluations were conducted separately for the seven fields: Plaintiff (PA), Defendant (PP), Key Issues (PR), Facts (F), Legal Theory (T), Claims (P), and Issues for Decision (Q). TOT is the sum of all pontuations for a given model.

Table 3. Average and Total Scores by Field and by Model

Modelo	PA	PP	PR	F	T	P	Q	TOT
gemini-2.5-pro-exp-03-25	3.89	3.77	3.55	3.53	3.39	3.52	3.51	25.16
gemini-2.0-flash	3.84	3.58	3.12	3.15	3.07	3.07	2.97	23.80
gemma-3-27b-it	3.49	3.57	1.89	1.97	2.04	2.00	2.05	16.01
gemma-3-1b-it	3.55	3.56	1.58	1.51	1.59	1.72	1.60	15.11

4.2. Overall Performance

The results demonstrate a consistent superiority of the Gemini models over the Gemma family, particularly in tasks demanding higher levels of contextual reasoning and legal abstraction. This trend is especially evident in the semantic fields of FACTS (F), LEGAL_THEORY (T), and ISSUES_FOR_DECISION (Q), where only the Gemini models consistently achieved average ratings above level 3, according to the evaluation protocol.

The gemini-2.5-pro-exp-03-25 model outperformed all others across nearly all fields, attaining scores that reflect not only operational relevance but also implementation readiness in complex fields (e.g., 3.53 in F, 3.39 in T). In contrast, the gemma-3-1b-it model failed to exceed a score of 2.0 in the most demanding tasks, with outputs frequently classified as imprecise, incomplete, or irrelevant—thereby falling below the operational threshold defined by the evaluation scale.

While the gemma-3-27b-it model showed modest gains over its 1B counterpart, its performance remained insufficient to meet the requirements of complex information extraction without domain-specific training. These findings reinforce that mere increases in model size are not sufficient to ensure semantic fidelity and task adequacy in specialized legal domains.

Furthermore, the comparative analysis suggests that model capability must be assessed not only in terms of aggregate scores, but also through qualitative distinctions in output quality. For instance, the Gemini models exhibited fewer instances of incoherent or noisy responses—an issue recurrent in Gemma models—highlighting their robustness in maintaining semantic integrity.

Finally, although the gemini-2.0-flash model exhibited slightly lower scores than the 2.5-pro variant, its performance remained within the threshold of operational usefulness (level 3) across most fields, which may justify its application in scenarios where computational efficiency is prioritized over maximal precision. While average scores offer useful exploratory insights, future work should incorporate statistical significance testing (e.g., Kruskal-Wallis or Friedman tests) to confirm whether observed performance differences are robust.

4.3. Performance by Specific Field

Comparative results for model performance on specific field can be seen on Figure 3. It was observed that the fields PLAINTIFF (PA) and DEFENDANT (PP) presented high

average scores across all evaluated models. Such performance can be attributed to the relatively objective nature of the task of identifying the parties involved, given that this information is generally explicitly stated in the initial sections of the petitions.

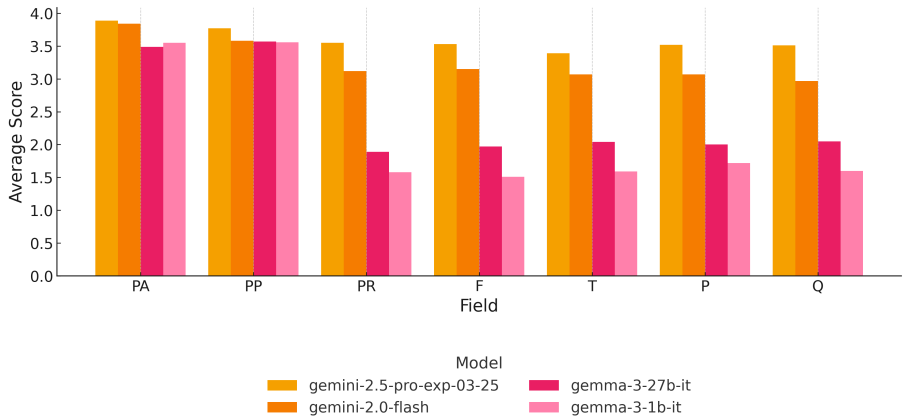

Figure 3. Comparative Model Performance on Specific Fields.

Fields that demand greater contextual understanding, such as FACTS and LEGAL_THEORY, posed substantial challenges for smaller models, which often failed to reconstruct coherent narratives or synthesize dispersed legal arguments. Consequently, outputs from models like gemma-3-1b-it and gemma-3-27b-it were frequently fragmented or inconsistent. In one case, a contractual rescission request was reduced to the phrase "right to cancellation," omitting the specific legal foundation in the Consumer Defense Code and undermining its practical applicability. Similar limitations were observed in the KEY_ISSUES field, where gemini-2.5-pro achieved an average score above 3.5, while both Gemma models remained below 2.0, reflecting a high level of ambiguity and missing information. In the CLAIMS field, although performance was generally better, low-capacity models continued to omit or oversimplify elements. For instance, ancillary requests such as "award of attorney's fees" were extracted in isolation, without reference to the main claim, resulting in operational inconsistencies. The ISSUES_FOR_DECISION field followed a comparable pattern. While high-capacity models like gemini-2.5-pro-exp-03-25 successfully identified the legal questions to be adjudicated, smaller models often generated vague or generic statements. One example, "The court shall decide according to the law," was rated at the lowest relevance level, as it failed to add value for understanding or automating judicial decision-making.

Additionally, the generation of content classified as "textual noise" was observed, primarily in the smaller models. This phenomenon included undue repetitions, irrelevant automated introductions, and excerpts disconnected from the procedural narrative. Such occurrences significantly compromise the reliability of automation systems that rely on these extractions, especially when potentially harmful information is inadvertently generated.

Smaller models frequently produced textual noise, including redundant phrases, irrelevant automated introductions, and excerpts disconnected from the procedural narrative. These issues compromise the reliability of automated systems, especially when the output contains misleading or potentially harmful content. Such findings emphasize

that achieving extraction quality at levels 3 and 4 requires models with strong contextual reasoning and, ideally, legal domain specialization. However, due to the operational costs of high-end models, a hybrid strategy may offer a more economical alternative. Simpler fields such as PA and PP can be handled by lightweight models like gemma-3-1b, while more complex fields may be assigned to gemini-2.0-flash, reserving gemini-2.5-pro for critical or high-precision use cases.

4.4. Performance by Jurisdiction Type

To analyze the impact of procedural jurisdiction on model performance, the results were stratified between petitions originating from the Special Civil Courts (JE Cível) and those from the Ordinary Civil Courts (Cível). Table 4 presents the average scores obtained in each semantic field, considering only the Gemini models separately.

Table 4. Average and Total Scores by Field, by Model and Jurisdiction Type

Model	Jurisdiction	PA	PP	PR	F	T	P	Q	TOT
gemini-2.0-flash	J.E. Cível	3.81	3.58	3.12	3.12	2.96	3.09	3.00	22.68
gemini-2.0-flash	Cível	3.88	3.58	3.12	3.19	3.21	3.05	2.93	22.95
gemini-2.5-pro	J.E. Cível	3.96	3.72	3.63	3.54	3.32	3.61	3.51	25.30
gemini-2.5-pro	Cível	3.79	3.84	3.44	3.51	3.49	3.40	3.51	24.98

It was found that the variations between jurisdictions were subtle, with differences of less than 0.3 points in nearly all fields. Nevertheless, certain relevant trends were observed and warrant specific attention. Regarding gemini-2.5-pro model, higher performance was observed for petitions from the Special Civil Courts in the fields of PA, PR, P, and Q, while petitions from the Ordinary Civil Courts showed a slight advantage in the fields of PP and T. This outcome may be explained by the greater degree of standardization and conciseness in documents from the Special Civil Courts, which tends to favor the extraction of direct and structured data. For example, the average score of 3.96 in the PA field for JE Cível suggests that identification of the plaintiff in these documents tends to be clearer and more explicit, facilitating its classification as ”highly relevant” (level 4 in the protocol). In contrast, the higher score for T in petitions from the Ordinary Civil Courts (3.49 versus 3.32) may be associated with the more developed argumentative structure found in these documents, which, although more complex, provide richer textual input for high-capacity inference models, such as gemini-2.5-pro, to correctly reason on it.

For the gemini-2.0-flash model, the differences between jurisdictions were even more subtle. A slight advantage was observed for Ordinary Civil petitions in the fields of F (3.19 versus 3.12) and T (3.21 versus 2.96), suggesting that, even for models with lower contextual capacity, more elaborated petitions enable the extraction of more legally relevant information. However, the model showed slightly lower performance in the Q field for the Ordinary Civil Courts (2.93), indicating greater difficulty in locating such information in longer and more sophisticated petitions.

In general terms, the results indicate that model performance is sensitive to the degree of structure and argumentative density of the analyzed documents. While texts from the Special Civil Courts favor the extraction of more objective elements (e.g., parties, requests), documents from the Ordinary Civil Courts offer better conditions for identifying

more complex elements (e.g., legal theses, facts), provided the model possesses adequate interpretative capacity.

5. Conclusion

This study investigated the feasibility of using LLMs for the task of automatically extracting structured information from initial petitions in the Brazilian legal context. To this end, a methodology was proposed based on advanced prompt engineering techniques combined with the use of state-of-the-art language models with standardized outputs in JSON format. Particularly, this work contributed by showing the feasibility of LLM-based extraction of seven key fields essential for judicial processing: Plaintiff, Defendant, Key issues, Facts, Legal Thesis, Requests, and Issues for Decision. Furthermore, the standardization of outputs in JSON format offers a promising pathway for future integration with case management systems, potentially contributing to the efficiency of various stages of the judicial workflow.

The empirical results indicated superior performance by the Gemini models, particularly in the extraction of information requiring greater contextual inference capabilities, such as the "Facts" and "Legal Theory" fields. Nevertheless, the competitive performance of these smaller models in more objective fields such as "Plaintiff" and "Defendant" suggests the potential for developing hybrid approaches that combine different models to balance the cost-effectiveness of the solutions.

As future work, it is proposed to expand the evaluation corpus to include different branches of Law, to refine prompt engineering strategies based on human feedback, and to incorporate automatic metrics that complement qualitative assessments. It is also intended to investigate the impact of supervised fine-tuning in specific legal domains, with the aim of increasing the robustness, consistency, and reliability of the generated extractions, as well as to test the proposed approach on next-generation language models in order to validate its scalability and adaptability to new architectures.

References

Adhikary, S., Sen, P., Roy, D., and Ghosh, K. (2024). A case study for automated attribute extraction from legal documents using large language models. *Artificial Intelligence and Law*, pages 1–22.

Breton, J., Billami, M. M., Chevalier, M., Nguyen, H. T., Satoh, K., Trojahn, C., and Zin, M. M. (2025). Leveraging llms for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*, pages 1–27.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. (2021). Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Conselho Nacional de Justiça (2025). Atos normativos - cnj. Last accessed 2025/04/13.

de Aquino, I. V., dos Santos, M. M., Dorneles, C. F., and Carvalho, J. T. (2024). Extracting information from brazilian legal documents with retrieval augmented generation. In *Proceedings of the Brazilian Symposium on Databases (SBBD)*, pages 280–287. SBC.

Google DeepMind (2025). Gemini continues to improve its ability to reason. Published March 2025, last accessed 2025/04/13.

Hussain, A. S. and Thomas, A. (2024). Large language models for judicial entity extraction: A comparative study. *arXiv preprint arXiv:2407.05786*.

Jayatilleke, N., Weerasinghe, R., and Senanayake, N. (2024). Advancements in natural language processing for automatic text summarization. In *Proceedings of the 2024 4th International Conference on Computer Systems (ICCS)*, pages 74–84. IEEE.

LM Studio (2025). Lm studio – run local llms, no api keys required. Last accessed 2025/04/13.

Ma, Z., Chen, A. R., Kim, D. J., Chen, T.-H., and Wang, S. (2024). Llmparser: An exploratory study on using large language models for log parsing. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Ngo, H. Q., Nguyen, H. D., and Le-Khac, N.-A. (2023). Building legal knowledge map repository with nlp toolkits. In *Proceedings of the 12th Conference on Information Technology and Its Applications (CITA 2023)*, volume 734 of *Lecture Notes in Networks and Systems*, pages 25–36. Springer.

Presidência da República (2018). Lei nº 13.709, de 14 de agosto de 2018 — lei geral de proteção de dados pessoais (lgpd). Last accessed 2025/04/13.

Sakiyama, K., Montanari, R., Junior, R. M., Nogueira, R., and Romero, R. A. F. (2023). Exploring text decoding methods for portuguese legal text generation. In *Proceedings of the Brazilian Conference on Intelligent Systems*, pages 63–77. Springer.

Souza, F., Souza, R., Neves, M., and Moreira, V. (2023). Legalbert-pt: Pre-trained language model for portuguese legal text. In *Proceedings of the Brazilian Conference on Intelligent Systems*. Springer.

Supriyono, W., Wibawa, A. P., Suyono, and Kurniawan, F. (2024). A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*, 7:100070.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., and et.al., T. M. (2025). Gemma 3 technical report.

Wan, Z., Zhang, Y., Wang, Y., Cheng, F., and Kurohashi, S. (2024). Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on chinese legal domain. *arXiv preprint arXiv:2310.03328*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, page 24824.

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., and Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1029–1046.

Zin, M. M., Satoh, K., and Borges, G. (2024). Leveraging llm for identification and extraction of normative statements. In *Legal Knowledge and Information Systems*, pages 215–225. IOS Press.