



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

João Pedro Perez Resmer

**PARAMETER-EFFICIENT FINE-TUNING AND ALIGNMENT OF QUANTIZED LLMS
FOR STRICT SCHEMA LEGAL EXTRACTION**

Florianópolis
2026

João Pedro Perez Resmer

**PARAMETER-EFFICIENT FINE-TUNING AND ALIGNMENT OF QUANTIZED LLMS
FOR STRICT SCHEMA LEGAL EXTRACTION**

Trabalho de Conclusão de Curso submetido
ao Curso de Graduação em Ciência da
Computação do Centro Tecnológico da
Universidade Federal de Santa Catarina
para obtenção do título de Bacharel em
Ciência da Computação.

Orientador: Prof(a). Carina Friedrich Dorne-
les, Dr(a).

Coorientador: Vitoria Soares dos Santos

Florianópolis
2026

Ficha catalográfica para trabalhos acadêmicos

Resmer, João Pedro Perez
Parameter-Efficient Fine-Tuning and Alignment of
Quantized LLMs for Strict Schema Legal Extraction / João
Pedro Perez Resmer ; orientador, Carina Friedrich
Dorneles, coorientador, Vitoria Soares dos Santos, 2026.
54 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro
Tecnológico, Graduação em Ciências da Computação,
Florianópolis, 2026.

Inclui referências.

1. Ciências da Computação. 2. Ajuste Fino. 3. Grandes
Modelos de Linguagem. 4. Extração de Dados. 5. Improbidade
Administrativa. I. Dorneles, Carina Friedrich. II. dos
Santos, Vitoria Soares. III. Universidade Federal de Santa
Catarina. Graduação em Ciências da Computação. IV. Título.

João Pedro Perez Resmer

**PARAMETER-EFFICIENT FINE-TUNING AND ALIGNMENT OF QUANTIZED LLMS
FOR STRICT SCHEMA LEGAL EXTRACTION**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de “bacharel em Ciência da Computação” e aprovado em sua forma final pelo Curso de Graduação em Ciências da Computação.

Florianópolis, 12 de julho de 2026.

Profa. Lúcia Helena Martins Pacheco, Dra. Coordenadora do Curso

Banca Examinadora:

Prof(a). Carina Friedrich Dorneles, Dr(a).
Carina Friedrich Dorneles, Dr.
Orientadora

Vitoria Soares dos Santos.
Coorientador
Universidade Federal de Santa Catarina

Jerusa Marchi, Dr.
Avaliador
Universidade Federal de Santa Catarina

Eduardo Camilo Inacio, Dr.
Avaliador
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Aos meus pais, **Bertha e Luciano**, por me ensinarem quase tudo que eu sei e valorizo, por me apoiarem e me corrigirem nas minhas decisões, por me proporcionarem oportunidades incríveis e por sacrificarem tempo e esforço por mim todos os dias da minha vida.

Ao meu irmão, **Bernardo**, por me fazer companhia e me dar a oportunidade de te ver crescer e se tornar um amigo com quem eu posso contar.

Aos meus avós, **João Mário, Célia, Odeti e Guillermo**, por me apoiarem sempre e me ensinarem a valorizar a família e os amigos, por me fazerem sentir amado e acolhido.

A minha namorada, **Luíza**, por me apoiar e me motivar a ser uma pessoa melhor todos os dias, por me fazer sorrir, me confortar e me fazer sentir especial.

Aos meus tios, **Marcelo, Angela, Fábio, Karen, Marcelo e Lucas**, por cada pequeno ensinamento e gesto que fazem parte do que me tornei.

A minha madrinha, **Monalisa**, por todo o carinho ao longo dos anos.

A minha madrinha de crisma, **Andreia**, pelo acolhimento e o carinho que me deu.

Aos meus primos, **Milena, Juliana, Felipe, Mateus, Gabriel, Cecília, Cícero, Beatriz, Vicente, Uma, Mia, Rafael e Julia**, aos que cresceram comigo e aos que vieram depois, por me fazerem sentir parte, por me ensinarem convivência, carinho, amizade, respeito e frustração, e por serem parte da minha vida e da minha história.

Aos meus host parents, **Christie e Steve**, por me mostrarem um outro lado do mundo, me ensinarem a valorizar o que eu tenho e me acolher de coração aberto.

Aos meus host brothers, **Jonah e Jake**, por me fazerem sentir em casa e me ensinarem a ser mais paciente e compreensivo.

Aos meus amigos, **Gabriel, Elita, Felipe, Guilherme, SanCleir, Fábio, Pedro Augusto, Pedro, Lucas** e todos os outros que fizeram parte da minha vida, pelas risadas, pelas conversas, pelos conselhos, pelas brigas, pelas festas, pelos jogos, pelos estudos, pelas viagens e por cada momento que me fizeram sentir vivo.

As minhas orientadora e coorientadora, **Carina e Vitoria**, por me guiarem, me ensinarem e me apoiarem durante todo o processo de desenvolvimento do trabalho, por me ajudarem a crescer como pesquisador e como pessoa, e por me darem a oportunidade de trabalhar com vocês.

RESUMO

Grandes Modelos de Linguagem (*Large Language Models* ou *LLMs*) emergiram como poderosas ferramentas em múltiplas tarefas nos últimos anos. A habilidade desses modelos de aproximar-se da performance humana em uma variedade de áreas a partir de predição do próximo token tem atraído muito interesse dentro e fora da comunidade acadêmica. Dada a capacidade desses modelos de representar relações semânticas em espaços vetoriais de alta dimensionalidade e processar essas informações por meio de camadas de atenção, o presente trabalho propõe dois métodos de ajuste fino usando *Low-Rank Adaptation (LoRA)* e quantização para fazer uso das incríveis capacidades desses modelos para auxiliar uma tarefa que demanda grande esforço, extração de dados em documentos legais, enquanto, ao mesmo tempo, minimiza os requerimentos de *hardware* necessários para treinar um modelo de 4 bilhões de parâmetros. O modelo escolhido foi ajustado em conjuntos de treino de diferentes tamanhos (100, 200, 300, 400 e 470 exemplos) sob dois regimes diferentes. O primeiro regime consiste no uso apenas de Ajuste Fino Supervisionado, enquanto o segundo método combina o Ajuste Fino Supervisionado com o Treinamento por Reforço. Os modelos foram treinados sob diferentes ambientes e combinações de hiperparâmetros com o objetivo de encontrar uma configuração ótima e aprimorar os resultados. Por meio da validação, descobrimos que o melhor modelo foi produzido pelo regime de Ajuste Fino Supervisionado com 470 exemplos, produzindo resultados similares ou melhores do que a combinação de Ajuste Fino Supervisionado e Treinamento por Reforço em todas as métricas. Testes realizados no conjunto de teste revelaram que nosso modelo foi capaz de superar todos os modelos de tamanho similar, e até mesmo alcançar performance comparável a modelos maiores, demonstrando a eficácia da nossa abordagem em aproveitar os LLMs para tarefas de extração legal estruturada.

Palavras-chave: Grande Modelos de Linguagem. Low-Rank Adaptation. Aprendizado por Reforço.

ABSTRACT

Large Language Models (LLMs) have emerged as very powerful tools for multiple tasks over the past years. Their ability to approximate human performance in a variety of fields through next-token prediction has attracted much attention inside and outside the academic community. Given their ability to represent semantics in high-dimensional vector spaces using embedding models and process them using self-attention, this work proposes two fine-tuning approaches using Low-Rank Adaptation (LoRA) and quantization to leverage the great capabilities of LLMs for a labor-intensive task, data extraction from legal documents, while minimizing the hardware requirements for training a 4-billion-parameter model. The chosen model was fine-tuned on training sets of different sizes (100, 200, 300, 400, and 470 examples) under two different regimes, a pure Supervised Fine-Tuning (SFT) approach and a second approach combining Supervised Fine-Tuning with a reinforcement learning step, and under different sets of hyperparameters, in order to find the optimal setup and improve the results. Through validation we found that the best model was produced by the pure SFT approach with 470 examples, producing similar or better results than the SFT and RL combination across all metrics. While testing on the held-out test set against our baseline models revealed that our model was able to outperform all models of similar size, and even achieve performance comparable to larger models, demonstrating the effectiveness of our approach in leveraging LLMs for structured legal extraction tasks.

Keywords: Large Language Models. Low-Rank Adaptation. Reinforcement Learning.

CONTENTS

1	INTRODUCTION	9
1.1	Research Questions	10
1.2	Objectives	10
1.2.1	General Objective	10
1.2.2	Specific Objectives	11
1.3	Main Contributions	12
1.4	Organization	13
2	BACKGROUND	14
2.1	Natural Language Processing	14
2.1.1	BoW and n-grams	14
2.1.2	Embeddings	15
2.1.3	Transformers	15
2.1.4	Sentence Embeddings and Cosine Similarity	16
2.2	Large Language Models	17
2.2.1	Training Large Models	17
2.2.2	In-Context Learning	18
2.2.3	Quantization	18
2.2.4	LoRA	19
2.3	Adam Optimizer	19
2.3.1	AdamW	19
2.4	Reinforcement Learning	20
2.4.1	Reinforcement Learning from Human Feedback (RLHF)	20
2.4.2	Proximal Policy Optimization (PPO)	20
2.4.3	Dynamic Sampling Policy Optimization (DAPO)	21
2.4.4	Reward Shaping for Structured Data Extraction	22
3	RELATED WORK	23
3.1	Domain-Specific LLMs for Law	23
3.2	Local Language Adaptation for Legal Text	23
3.3	LLMs for Document Information Extraction	24
3.4	Multi-Task Instruction Tuning for Unified Extraction	24
3.5	Structured Generation and Constrained Decoding	24
3.6	Synthesis and Research Gaps	25
4	METHOD	27

4.1	Generative AI usage	27
4.2	General View of the Proposed Approach	27
4.3	Task Formulation	28
4.4	Document Structure	29
4.5	Dataset Annotation	32
4.6	Setup	33
4.7	Quantization and Model Initialization	33
4.8	Low-Rank Adaptation (LoRA) Configuration	33
4.9	Supervised Fine Tuning Description	34
4.9.1	Custom Loss Function	34
4.10	Reinforcement Learning (Two-Stage Approach)	36
4.10.1	DAPO Configuration and Training Dynamics	36
4.10.2	Programmatic Reward Shaping	36
4.11	Experiment Design	37
4.12	Prompting	39
4.13	Validation	39
5	RESULTS	41
5.1	Comparison of The Two Tuning Approaches	41
5.2	Test Set Evaluation	41
5.2.1	Per-Field Granular Analysis	44
6	CONCLUSION AND FUTURE WORK	46
	BIBLIOGRAPHY	48
A	SBC PAPER	51
B	SOURCE CODE AND DATA	66

1 INTRODUCTION

As digitization spreads across various fields worldwide, Brazil is no exception. In fact, the country currently has an extensive array of digitized documents and procedures available. A notable example is the national ID, which, since 2022, has featured a valid online version accessible to anyone who has had their document issued since then. The judicial system has followed this push toward digitized procedures. This shift has significantly enlarged the production and storage of legal documents, making them a vast source of information for legal professionals such as lawyers, judges, and prosecutors. However, this creates a new challenge: extracting useful information from a massive database of documents that are semantically complex and extremely diverse in length, content, context, and structure to then formatting it into a schema adherent json object is a highly difficult and labor-intensive task.

Specifically, between 2020 and 2024, Brazil registered over 205,000 procedures related solely to bidding processes. Each of these procedures requires multiple documents to be issued, creating a massive influx of information into the Brazilian government's systems. Consequently, manually reviewing these documents to search for useful information applicable to a specific subset of judicial procedures becomes an extremely time-consuming task.

Given these challenges, it is natural to turn to computational solutions. Natural Language Processing (NLP) has long been an area of interest, and various methods have been developed to extract specific sequences of characters from text. The scientific community has utilized many different approaches, ranging from pattern matching (CHENG et al., 2009) to regular expression-based methods, achieving varying levels of performance. While consistent and structured data with clear formatting rules can be efficiently retrieved, extracting relevant information from diverse, unstructured excerpts has remained challenging. This remained the case until the introduction of the Transformer architecture (VASWANI et al., 2017), which significantly advanced the field of natural language processing.

This new architecture allowed deep learning models to effectively process semantics in text embeddings and achieve unprecedented text generation capabilities. However, certain difficulties persisted. As noted by (HUANG et al., 2023), these models often lack domain-specific knowledge and struggle to apply general knowledge to solve domain-related tasks. Despite these challenges, recent studies have demonstrated remarkable performance in information extraction tasks. For instance, (WAN et al., 2024) proposed a zero-shot, three-step pipeline capable of enhancing the performance

of GPT-4 on Chinese legal documents, while (GODOY, 2025) employed fine-tuning techniques to improve performance on a more diverse dataset.

The present work focuses on addressing these gaps by training a model to achieve state-of-the-art performance on a legal data extraction task. While commercial, cloud-based Large Language Models are powerful, their application in the legal domain is often restricted by strict document confidentiality requirements and prohibitive inference costs at scale. Therefore, this research emphasizes a local processing approach. Using quantization and Low-Rank Adaptation (LoRA), we locally train a 4-billion-parameter model (Gemma 3). The model is fine-tuned on a custom dataset consisting of 100 legal documents (50 hand-annotated, and 50 automatically annotated with Llama 3.3 using a 10-shot prompting strategy and subsequently human-revised) related to administrative impropriety.

1.1 RESEARCH QUESTIONS

The task of extracting structured data from legal documents is highly complex, requiring the model to comprehend intricate legal semantics and output a strictly schema-adherent JSON object. To address this challenge, we formulated the following research questions:

1. **RQ1:** How does the scaling of the fine-tuning dataset size (from 100 to 470 examples) impact the model's ability to minimize formatting errors and lexical hallucinations in structured legal data extraction?
2. **RQ2:** How does a hybrid training approach combining Supervised Fine-Tuning (SFT) with Reinforcement Learning (RL) compare to standard SFT in terms of structural schema adherence and domain adaptation?
3. **RQ3:** Can a quantized, LoRA-adapted 4B-parameter model trained locally match or exceed the information extraction performance of baseline legal NLP approaches while satisfying data confidentiality constraints?

1.2 OBJECTIVES

This section outlines the objectives of the present work.

1.2.1 General Objective

The general objective of this work is to fine-tune a large language model using quantization and LoRA, guided by a custom loss function that tolerates variations in wording as long as the semantic meaning is preserved. This fine-tuned model must be

capable of effectively extracting relevant information from legal documents concerning administrative impropriety, outputting the extracted data in a structured JSON format.

1.2.2 Specific Objectives

1. **Compile and preprocess a specialized dataset:** Construct a custom dataset from 100 legal documents on administrative impropriety, segmenting them to adhere to token limits and pairing them with their corresponding structured JSON outputs.
2. **Implement a parameter-efficient training pipeline:** Set up the foundational training environment using quantization and Low-Rank Adaptation (LoRA) to fine-tune a 4-billion-parameter language model locally.
3. **Develop and integrate a semantic-based loss function:** Formulate and evaluate a custom loss function based on cosine similarity to prioritize semantic accuracy over strict lexical matching during the fine-tuning process.
4. **Apply Reinforcement Learning:** Implement an alternative training approach combining traditional Supervised Fine-Tuning (SFT) with Reinforcement Learning (RL) to analyze its impact on the model's domain adaptation.
5. **Optimize training hyperparameters:** Conduct systematic testing to identify an optimal or near-optimal set of hyperparameters for the proposed training configurations.
6. **Evaluate and compare model performance:** Assess the performance of the resulting models against established baselines using quantitative metrics to validate the effectiveness of the proposed extraction pipeline.

With these objectives, the present work aims to advance the field of legal information extraction by demonstrating that smaller, locally trained models can achieve state-of-the-art performance in highly specialized domains, significantly reducing the hardware barriers typically associated with Large Language Models. This must be achieved by ensuring that the model can effectively extract relevant information from legal documents while adhering to strict schema requirements and maintaining data confidentiality.

Specifically, the structured output must accurately capture information across three core fronts: *dados da licitação* (bidding data), *dados do processo* (process data), and *dados dos requeridos* (defendants' data). To comply with a 2048-token limit, each document was segmented into multiple instances, with each instance comprising a

document excerpt paired with its corresponding structured JSON object. As a result, the annotation process produced over 570 distinct examples. To ensure rigorous evaluation, 100 of these examples were reserved exclusively as a test set, while the remainder constituted the training data. Additionally, we employ a custom loss function during fine-tuning that leverages cosine similarity to more accurately score outputs that capture the correct underlying semantic information. In order to better understand the model's capabilities to adapt to the task under different training settings, we also employ a second approach combining Supervised Fine-Tuning with Reinforcement Learning. Dynamic Sampling Policy Optimization (DAPO) with a programmatic reward to directly optimize the policy against structural adherence and semantic precision, subsequently comparing the results in order to produce a final model (LegalExtract470).

1.3 MAIN CONTRIBUTIONS

This section presents the main contributions of this work.

1. **A Curated Domain-Specific Dataset:** We introduce a novel dataset of 100 legal documents on administrative impropriety, annotated with structured JSON outputs. This dataset serves as a benchmark for evaluating information extraction models in the Brazilian legal domain.
2. **A Parameter-Efficient Fine-Tuning Pipeline:** We propose a fine-tuning pipeline that leverages quantization and Low-Rank Adaptation (LoRA) to adapt a 4-billion-parameter language model for structured data extraction tasks, demonstrating that smaller, locally trained models can achieve state-of-the-art performance in highly specialized domains.
3. **A Custom Semantic Loss Function:** We develop a novel loss function that prioritizes semantic accuracy over strict lexical matching, enabling the model to better capture the underlying meaning of legal text and produce more accurate structured outputs.
4. **A Comparative Analysis of Training Approaches:** We conduct a comprehensive evaluation of Supervised Fine-Tuning (SFT) and a hybrid approach combining SFT with Reinforcement Learning (RL), providing insights into the trade-offs and benefits of each method in the context of structured legal data extraction.
5. **A Local Model for Legal Information Extraction:** We demonstrate the feasibility of deploying a locally trained, quantized 4-billion-parameter model for legal information extraction, addressing data confidentiality concerns and reducing the computational resources required for inference. While obtaining state-of-the-art

performance, this model significantly reduces the hardware barriers typically associated with Large Language Models, making it accessible for resource-constrained environments.

1.4 ORGANIZATION

The present document is structured as follows. Chapter 1 serves as an introduction to the present work and an overview of text mining and its importance. Chapter 2 explores important background concepts, such as Natural Language Processing, Large Language Models, and other related topics. Chapter 3 covers important foundational work. Chapter 4 presents the proposed approach for fine-tuning. Describes the experiments and metrics used for validation and testing. Chapter 6 is a brief discussion about the validation process and its results. Chapter 5 details the experimental results conducted for testing the final model and a comparison to baseline models. Finally, Chapter 6 presents the conclusion and synthesis of the work done.

2 BACKGROUND

This section presents fundamental theory related to the presented work theme.

2.1 NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence focused on the study of language. Its goal is to leverage techniques that allow automated human-like text generation and comprehension (JURAFSKY; MARTIN, 2024). Throughout the years, this area of study has introduced many interesting technologies, from purely statistical ones, such as Bag-of-Words (BoW) (SALTON; WONG; YANG, 1975; TURNEY; PANTEL, 2010), to the introduction of embedding models. Such models enable the representation of text semantics through high dimensional vector spaces popularized by word2vec (MIKOLOV et al., 2013). More recently, the field has been revolutionized by the adoption of attention mechanisms and the Transformer architecture (VASWANI et al., 2017).

2.1.1 BoW and n-grams

The Bag-of-words (BoW) (SALTON; WONG; YANG, 1975; TURNEY; PANTEL, 2010) model is a purely statistical method for modeling text. It consists of counting the frequency of each word as it appears in the text. Formally, it is representing text as a set of tuples, each constituted of a word that appears on the original document and an integer, the number of times that word appears on the document. It is efficient and, when coupled with machine learning or statistical analysis techniques, allows us to extract meaningful information for natural language.

However, Bag-of-Words fails to represent a crucial component of text: context. Sentences gain meaning through the combination of words in context with one another, a simple shift of replacement of a single word could entirely change the meaning of a sentence. With that in mind, another way of statistically modeling text are n -grams (SHANNON, 1951). The n -grams are sequences of n tokens (words, syllables or characters depending on granularity). This model allows us to collect statistically significant data of words in context and even next token prediction exclusively by calculating the probability of that sequence of words under the observed distribution (represented in Equation 2.1). Nevertheless, this model also carries limitations: increasing n exponentially grows the number of possible n -grams, which makes data collection and probability estimation difficult. It becomes computationally expensive to count and store all n -grams and significantly harder to avoid the zero-frequency (data sparsity) problem.

$$P(w_1, w_2, \dots, w_m) \approx \prod P(w_i | w_{i-(n-1)}, \dots, w_{i-1}) \quad (2.1)$$

2.1.2 Embeddings

In embedding techniques, instead of counting discrete units of text, tokens are mapped into a continuous vector space. The goal is to achieve a representation that maintains semantic properties by representing a such properties as numerical values in a vector. That is achieved by training a neural network (NN) in a task that induces it to produce this representation. For instance, the Word2vec model (MIKOLOV et al., 2013) achieves this by training a neural network to identify whether or not two words are contextually correlated. From these representations, it is possible to extract information such as the mathematical similarity between vectors. This concept is also used in more complex NLP models, including the Transformer architecture.

2.1.3 Transformers

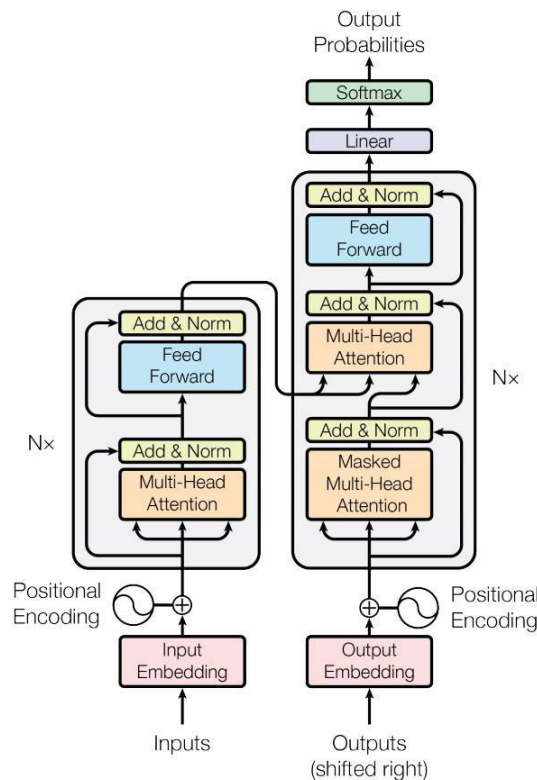
The introduction of the Transformer architecture (VASWANI et al., 2017) marks a paradigm shift in NLP and modern day artificial intelligence. The proposed architecture presents an encoder decoder base, but relies on a novel processing strategy Multi-head Self Attention. Attention, as defined in Equation 2.2, is a function that maps a query (Q) and a key-value (K, V) pair to an output. That is done by matrix multiplication, scaling and a softmax function. This output represents how relevant the query is to the key-value pair, calculated via matrix multiplication, scaling, and a softmax function. Multi-Head Attention (Equation 2.3) consists of projecting queries, keys, and values h times, calculating attention in each of these projection heads, concatenating the results, and projecting them a final time to produce the output.

$$Attention(Q, V, K) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2.2)$$

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \dots, head_h)W^O, \\ \text{where } head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (2.3)$$

As seen in Figure 1 the Transformer architecture uses self attention mechanisms both for the encoder and the decoder. However, there is an important distinction in the decoder: during generation previous outputs are shifted right and then passed through a modified self attention layer. This modified layer calculates a masked attention that guarantees new tokens are influenced only by previous tokens. The output of this layer is then used to calculate cross-attention, where queries (Q) come from the previous

Figure 1 – Transformer Architecture



decoder layer, and keys and values (K, V) are projected from the latent encoding produced by the encoder.

2.1.4 Sentence Embeddings and Cosine Similarity

While word-level embedding models focus on capturing semantic relationships between discrete tokens, real-world information extraction often demands the evaluation of complete textual blocks, such as sentences, paragraphs, or structured JSON representations. To bridge this gap, sentence embedding frameworks leverage transformer-based encoder architectures to map entire sequences of variable lengths into fixed-dimensional dense vectors (REIMERS; GUREVYCH, 2019). These models are typically trained using contrastive learning objectives, ensuring that semantically similar sentences are mapped to adjacent coordinates within the high-dimensional vector space.

To quantitatively evaluate the proximity between these dense representations, a vector distance or similarity metric must be employed. In high-dimensional semantic spaces, *Cosine Similarity* serves as the standard mathematical formulation, as it measures the cosine of the angle between two non-zero vectors regardless of their magnitude. Formally, given a predicted vector $\mathbf{A}$ and a target ground-truth vector $\mathbf{B}$ in an n -dimensional space, the cosine similarity is defined as:

$$\text{CosineSimilarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (2.4)$$

The resulting value is bounded between $[-1, 1]$, where 1 indicates identical directional alignment (perfect semantic convergence) and -1 represents total opposition. In the context of structured legal data extraction, this metric provides a continuous, differentiable signal that prioritizes underlying semantic equivalence over strict, rigid lexical matching.

2.2 LARGE LANGUAGE MODELS

While establishing an exact threshold for what constitutes a "Large" Language Model (LLM) is subjective, the foundational concept is well-defined: LLMs are highly scaled Transformer-based architectures characterized by billions of learnable parameters (ZHAO et al., 2023). They are trained on massive, diverse datasets of natural language and code, allowing them to excel at a wide variety of NLP and general reasoning tasks. Prominent examples include OpenAI's GPT family (such as GPT-3 and GPT-4) (OPENAI, 2023), Google's Gemini (TEAM et al., 2023), and Meta's open-weights LLaMA series (TOUVRON et al., 2023).

Furthermore, the rapid evolution of open-source and open-weights foundational architectures, such as Meta's LLaMA series (DUBEY et al., 2024), Google's Gemma (Gemma Team et al., 2024), and Alibaba's Qwen models (YANG et al., 2024), has marked a significant paradigm shift in the field. By providing public access to high-performance model weights, these architectures have democratized advanced natural language capabilities. Crucially, they enable organizations to execute local inference and perform parameter-efficient fine-tuning on consumer-grade or specialized hardware, effectively bypassing the strict data confidentiality constraints and prohibitive API licensing costs traditionally associated with proprietary cloud-based models.

2.2.1 Training Large Models

Training LLMs is a difficult task, and its challenges stem mainly from the nature of the models. In order to create a model that seems to understand human language, these massive artificial intelligence models are trained on unprecedented volumes of data, which implies massive computational costs. Consequently, the field is constantly challenged to make this training process less expensive, more stable, and capable of generating state-of-the-art results.

Another product of this necessity for extensive training, the process is typically split of training into two main stages. First, the model is trained on said massive dataset of natural language. In this phase, known as *pre-training*, the objective is next-token

prediction to learn grammar, facts, and reasoning abilities. However, a model that strictly completes text by sampling from its vocabulary distribution is not always fit for practical applications, as it may lack instruction-following capabilities or format adherence (such as outputting strict JSON schemas).

Therefore, a second stage called *fine-tuning* is applied. The model is adapted for a more specific setting or task while preserving the foundational knowledge acquired during pre-training. This stage is substantially less computationally expensive, especially with the advent of Parameter-Efficient Fine-Tuning (PEFT) techniques designed to optimize pre-trained models efficiently. Quantization and Low-Rank Adaptation (LoRA) (HU et al., 2021) are prime examples of such optimizations.

2.2.2 In-Context Learning

In-Context Learning (ICL) is a powerful capability of Large Language Models to adapt to a novel task simply by conditioning on a prompt containing a few demonstrations (few-shot prompting), without requiring any updates to the model's underlying weights. While fine-tuning techniques explicitly alter the model's parameters to optimize for a specific task, ICL leverages the knowledge acquired during pre-training to infer the task pattern directly from the provided text context.

The underlying mechanism of how Transformers achieve this phenomenon has been the subject of extensive research. In 2023, Google researchers (OSWALD et al., 2023) demonstrated that the forward pass of a Transformer during ICL implicitly performs gradient descent. According to their research, when a Transformer is presented with a sequence of input-output demonstrations in a prompt, its linear self-attention layers effectively compute a meta-optimization step.

By attending to the provided examples, the self-attention mechanism constructs an internal linear model, calculates an implicit weight update (mathematically analogous to standard gradient descent) and applies it internally to predict the final target token. This demonstration provides context to the significant improvements that few-shot prompting brings to generation.

2.2.3 Quantization

Quantization is a technique used to reduce the memory demand of LLMs by lowering the precision of their weights (GHOLAMI et al., 2021). In standard pre-training, weights are stored with high precision (e.g., 32-bit or 16-bit floating-point), as this directly impacts on the models capabilities of representing knowledge. Quantization trades a marginal amount of precision-and thus a fraction of performance-for significant reductions in required computational resources. Enabling the operation and fine-tuning

of Large Language Models using drastically less VRAM, energy, and time (DETTMERS et al., 2022), by using 8-bit or 4-bit representations.

2.2.4 LoRA

Low-Rank Adaptation (LoRA) (HU et al., 2021) is a training technique used to increase fine-tuning efficiency. Instead of updating the entire dense weight matrix according to the gradient, it freezes the pre-trained model and updates only trainable rank decomposition matrices (the adapters) for targeted layers. This is achieved by representing the weight update as shown in Equation 2.5. By keeping the main weights (W) frozen and only updating the smaller matrices (B and A) of a low rank r , LoRA saves computational resources, memory, and time. Furthermore, LoRA allows targeted training, as the layers without injected adapters remain frozen.

$$W' = W + BA, \text{ where } B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k} \quad (2.5)$$

2.3 ADAM OPTIMIZER

The Adam (Adaptive Moment Estimation) optimizer (KINGMA; BA, 2014) is a deep learning training algorithm. The algorithm maintains exponentially decaying estimates of both first (mean) and second (uncentered variance) moments of the gradients to improve the stability and convergence speed during training. These estimates are updated as in Equations 2.7 and 2.8 respectively, where the gradient g_t is defined as in Equation 2.6. Notice that these definitions of gradient moment require bias correction as they tend to zero earlier in training. The weight update, then is defined in Equation 2.9 to enable adaptive learning rate.

$$g_t = \nabla_{\theta} f(\theta_t) \quad (2.6)$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2.7)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (2.8)$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \quad (2.9)$$

2.3.1 AdamW

In the regular Adam optimizer, weight decay acts as L2 regularization. However, applying L2 regularization directly to the adaptive gradient can lead to suboptimal weight decay. AdamW (LOSHCHILOV; HUTTER, 2017) fixes this by separating the decay from the gradient to make them independent (as seen in the $\eta \lambda \theta_t$ term in Equation 2.10).

Modern LLM fine-tuning often leverages optimized variants of AdamW. For example, *Paged AdamW* stores optimizer states in memory pages, enabling the system to offload states to system RAM when VRAM is constrained. Furthermore, *8-bit AdamW* quantizes the optimizer’s moment estimates to heavily reduce storage requirements at the cost of minimal precision loss.

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} - \eta \lambda \theta_t \quad (2.10)$$

2.4 REINFORCEMENT LEARNING

Reinforcement Learning (RL) is a paradigm of machine learning in which an *agent* learns to make decisions by performing *actions* within an *environment* to maximize a cumulative *reward* (SUTTON; BARTO, 2018). Formally, this is modeled as a Markov Decision Process (MDP), defined by a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, transition probabilities $\mathcal{P}$, and a reward function $\mathcal{R}$. The goal is that the agent learns an optimal policy $\pi(a|s)$ —a probability distribution over actions given a state—that maximizes the expected cumulative reward over time.

While pre-training and Supervised Fine-Tuning (SFT) teach a Large Language Model vocabulary, grammar, and basic reasoning, these paradigms strictly rely on next-token prediction via cross-entropy loss. They do not inherently guarantee that the model will follow specific constraints, avoid hallucinations, or align with complex human preferences. To address this, RL is applied to language modeling by treating the LLM as the agent, the input prompt and previously generated tokens as the state, the next token generation as the action, and a scoring mechanism as the environment’s reward.

2.4.1 Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) (OUYANG et al., 2022) is a training methodology designed to align LLMs with desired behaviors. In this framework, a secondary model called the Reward Model (RM) is trained on datasets of human preferences to score the quality of the LLM’s outputs. Once the RM is trained, the primary LLM is optimized using RL algorithms to generate text that yields high reward scores. This technique was instrumental in the development of models like InstructGPT and modern conversational agents, as it heavily penalizes undesired outputs (such as toxic or factually incorrect text) that standard fine-tuning might otherwise permit.

2.4.2 Proximal Policy Optimization (PPO)

To update the LLM’s weights based on the reward signal, RLHF typically utilizes Proximal Policy Optimization (PPO) (SCHULMAN et al., 2017). Updating an LLM

using standard policy gradient methods can lead to destructively large parameter updates, causing the model to forget its pre-trained knowledge (catastrophic forgetting) or generate unnatural text.

PPO mitigates this by restricting how much the policy can change in a single update step. It achieves this by defining a probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ and optimizing a clipped surrogate objective function, as shown in Equation 2.11. By clipping the ratio between $[1 - \epsilon, 1 + \epsilon]$, PPO ensures that the new policy π_θ does not deviate excessively from the old policy $\pi_{\theta_{\text{old}}}$, leading to highly stable training.

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (2.11)$$

2.4.3 Dynamic Sampling Policy Optimization (DAPO)

While PPO is foundational, it requires maintaining multiple models simultaneously (actor, reference, reward, and value networks), making it highly computationally expensive. Recently, methods like Group Relative Policy Optimization (GRPO) (SHAO et al., 2024) gained traction for reasoning tasks by eliminating the need for a value network, relying instead on comparing multiple sampled outputs for the same prompt. However, standard group-based RL often suffers from training instability, entropy collapse, and inefficiencies in credit assignment for long sequences.

To overcome these scalability and stability issues, Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) (LIU et al., 2025) was introduced as a highly efficient framework for large-scale LLM reinforcement learning. DAPO builds upon base group optimization by introducing several critical techniques:

- **Dynamic Sampling:** During training, DAPO actively filters out prompts where all sampled outputs yield an identical reward (e.g., all completely correct or all completely wrong). By dropping samples that provide a zero-advantage signal, dynamic sampling drastically improves training efficiency and focuses computational resources only on informative data batches.
- **Clip-Higher Strategy:** Instead of a symmetrical clipping range (as seen in standard PPO), DAPO decouples the clipping bounds using separate low and high thresholds. This promotes policy diversity, encourages exploration, and prevents the entropy-collapsing phenomenon where the model becomes overly deterministic too early in training.
- **Token-Level Policy Gradient Loss:** Traditional methods average gradients at the sequence level, which heavily underweights the contribution of individual tokens in

long responses. DAPO computes gradients per token, ensuring stable learning and credit assignment even for exceptionally long generated outputs.

2.4.4 Reward Shaping for Structured Data Extraction

In the context of specific, high-stakes tasks—such as automated data extraction in the legal domain—the principles of RL and alignment become highly valuable. When extracting structured data from legal text, the output must not only be factually accurate (grounded strictly in the source text to prevent hallucinations) but also adhere to rigid syntactic structures, such as a predefined JSON schema.

Instead of relying solely on human feedback, the reward function can be programmatically shaped (often referred to as Reinforcement Learning from AI Feedback or Programmatic RL). In this setup, the environment automatically calculates the rewards based on structural validity. For instance, the agent receives a high positive reward for outputting valid JSON that matches the target schema, and a heavy negative penalty (negative reward) for syntax errors, missing keys, or hallucinated legal clauses. This application of RL ensures that the language model goes beyond text generation, effectively functioning as a reliable, constraint-abiding data extraction engine.

3 RELATED WORK

The task of extracting structured data from Brazilian legal documents using Large Language Models (LLMs) lies at the intersection of multiple areas of Natural Language Processing (NLP) research. In order to contextualize this work and its contributions, the present section reviews relevant literature across three primary domains. First, it explores the adaptation of LLMs to the legal domain. Second, it analyzes work on specialized models designed for document-level information extraction. Lastly, it focuses on instruction tuning for structured extraction. By examining these foundational works, this section establishes the current state-of-the-art on similar tasks and highlights research gaps.

3.1 DOMAIN-SPECIFIC LLMS FOR LAW

Domain specific adaptation for highly technical fields, such as law, have been a source of recent advancements. Amongst them, a notable contribution is SaulLM-7B (COLOMBO et al., 2024), a pioneering model designed for legal text. The model is an adaptation of Mistral 7B trained on an extensive legal corpus in order to comprehend legal vocabulary and reasoning. However, while SaulLM establishes the viability of such models for legal reasoning, its efforts were focused on on general legal comprehension in the English language. Additionally, the model’s training and evaluation is aimed towards more classical Legal NLP tasks, such as summarization. Consequently, the ability of smaller foundational models, such as Gemma-3:4B, to learn the syntactic and lexical nuances of Brazilian legal documents is still not widely explored.

3.2 LOCAL LANGUAGE ADAPTATION FOR LEGAL TEXT

While domain-specific models like SaulLM-7B demonstrate the power of continuous pre-training for legal language, they remain fundamentally limited by their focus on the English language and Anglo-Saxon legal systems. Within the Brazilian legal context, recent efforts have focused on adapting Large Language Models to Portuguese. For instance, the Sabiá family of models (PIRES et al., 2023) demonstrated that careful curation of Portuguese corpora yields competitive performance on local benchmarks without requiring massive scale.

However, a persistent research gap remains at the intersection of local language adaptation and structural extraction tasks. While models like Sabiá or Llama-BR excel at general text understanding and legal classification, they are rarely evaluated or optimized for generating strict, deeply nested structures—such as JSON schemas—under local

hardware constraints. This work addresses this limitation by taking Gemma 3, a highly efficient, multilingual small language model, and explicitly aligning it for structured extraction within the complex domain of Brazilian Administrative Misconduct actions.

3.3 LLMS FOR DOCUMENT INFORMATION EXTRACTION

Although domain adaptation may improve a model's capabilities for legal domain tasks in general, extracting structured information directly from unstructured documents requires specialized training. (GODOY, 2025) addressed this challenge with Extract-0, a specialized language model trained for document information extraction. Extract-0 highlights the efficacy of specializing language models for document parsing as an alternative to traditional pipeline-based NLP extraction. However, applying such methodologies to often ambiguous and more heterogeneous text, while producing highly structured, schema-dependent outputs—like strictly formatted JSONs required in legal tech workflows—presents distinct challenges that often require customized optimization strategies.

3.4 MULTI-TASK INSTRUCTION TUNING FOR UNIFIED EXTRACTION

To enable models to generalize across varied extraction tasks, researchers have increasingly relied on instruction tuning. InstructUIE, introduced by (WANG et al., 2023), is a comprehensive framework that utilizes multi-task instruction tuning to create a unified information extraction system. By framing various extraction tasks (such as entity, relation, and event extraction) as natural language generation tasks, InstructUIE proved that seq-to-seq models could handle complex, multi-faceted extraction instructions simultaneously. InstructUIE serves as a foundational proof-of-concept that instruction tuning is a highly effective method for teaching models to map unstructured text to structured outputs. However, their reliance on standard Supervised Fine-Tuning (SFT) utilizing standard Cross-Entropy (CE) loss may not be sufficient when rigid adherence to complex data schemas is strictly required.

3.5 STRUCTURED GENERATION AND CONSTRAINED DECODING

A critical challenge in leveraging LLMs for Information Extraction (IE) is ensuring that the generated output strictly adheres to a predefined schema (e.g., JSON). Traditional IE approaches rely heavily on fine-tuning to teach the model both the semantic extraction task and the syntactic rules of the target format. To mitigate formatting failures, recent literature has proposed mechanisms such as constrained decoding. A prominent framework is Outlines (WILLARD; LOUF, 2023), which enforces formal grammar con-

straints directly at the token sampling level during inference, effectively guaranteeing 100% syntactical validity without altering the model's internal weights.

In this study, we choose not to employ inference-time grammar constraints during the main evaluation phases. This decision allows us to isolate and measure the raw alignment capability of the SFT and DAPO training paradigms. By analyzing the model's unconstrained output, such as the 96% JSON validity rate observed, we can accurately assess how effectively the custom loss functions and reinforcement learning steps imbue the model with structural discipline, leaving constrained decoding frameworks like Outlines as an orthogonal, complementary optimization for production deployment.

3.6 SYNTHESIS AND RESEARCH GAPS

The reviewed literature establishes a strong foundation for this study: SaulLM (COLOMBO et al., 2024) validates the use of small-to-medium LLMs in the legal domain. Extract-0 (GODOY, 2025) demonstrates the value of models specialized for document extraction. And InstructUIE (WANG et al., 2023) illustrates the power of instruction tuning for seq-to-seq extraction.

However, synthesizing these works reveals a distinct gap in the literature. Most current approaches either focus on legal corpora or rely entirely on standard Supervised Fine-Tuning under the premise that it works. In the context of Brazilian legal documents where domain-specific terminology and text heterogeneity are predominant, standard SFT often yields hallucinated JSON keys or invalid formatting.

Furthermore, there is a lack of comparative studies investigating how advanced alignment techniques can solve these structural hallucination issues in smaller models (e.g., 4B parameters). This study directly addresses this gap by experimenting with Gemma-3:4B on a novel task setup. By comparing an SFT-only approach enhanced with a custom loss function against a hybrid approach, that utilizes standard CE loss for SFT, followed by Reinforcement Learning using Direct Alignment from Preferences Optimization (DAPO) and a custom reward function. This work introduces a novel methodology for enforcing strict schema constraints in non-English legal information extraction. Table 1 summarizes the key differences between the proposed approach and the most relevant related works.

Table 1 – Qualitative comparison of related works and the proposed approach.

Work / Model	Target Domain	Task Paradigm	Output Constraints	Con-	Optimization Method
<i>SaulLM-7B</i>	Legal (English)	General NLP & Text Generation	Unconstrained Free Text		Continuous Pre-training
<i>Extract-0</i>	General (English)	Information Extraction	Flat JSON Schema		Supervised Fine-Tuning
<i>InstructUIE</i>	General (Multilingual)	Unified Information Extraction	Sequence-to-Sequence		Instruction Tuning
<i>Sabiá</i>	General (Portuguese)	General Benchmark Tasks	Unconstrained Free Text		Pre-training Adaptation
This work	Legal (Brazilian Portuguese)	Information Extraction	Deeply Nested JSON		Custom Loss SFT and SFT + DAPO (RL)

4 METHOD

The following section details the proposed fine tuning approach. The first stage of fine tuning is loading the dataset and the model, then applying quantization, configuring LoRA, and finally training.

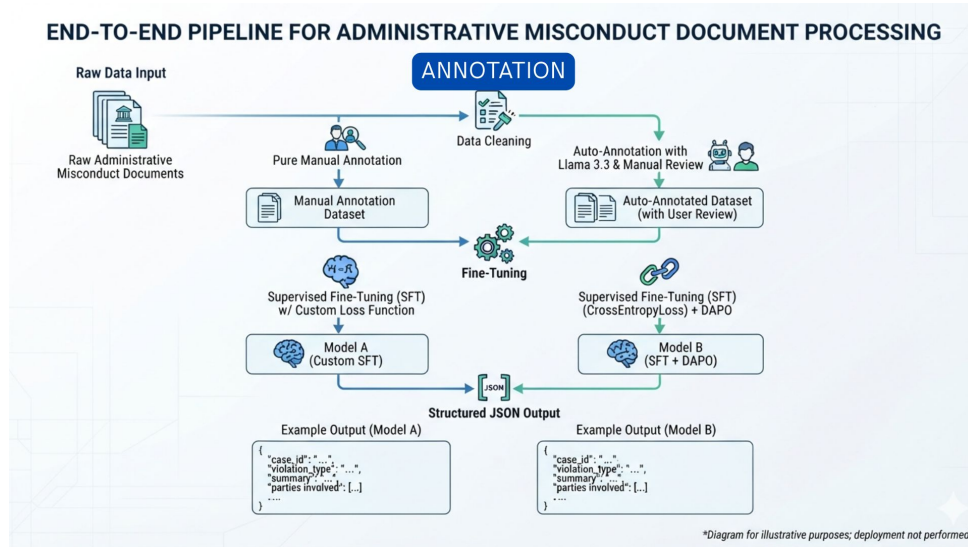
4.1 GENERATIVE AI USAGE

The use of Generative AI in this work was a constant, the models were used as an assisting tool at multiple stages of the project. During the annotation stage, Llama 3.3 was used to pre-annotate a subset of the dataset, which was then human verified and corrected. During the implementation stage, the models were used to assist in code generation, debugging and a source of information regarding the libraries used. During the training stage, the Gemma embedding model was used to generate text embeddings for the custom loss function, which combines cross-entropy with cosine similarity to reward semantically correct outputs even if they are lexically different from the ground truth. Finally, during evaluation, a lightweight Sentence Transformer (all-MiniLM-L6-v2) was used to compute semantic similarity scores between generated outputs and ground truth labels as part of the programmatic reward shaping in the reinforcement learning phase. During the writing of this report, the models were used to assist in drafting and editing the text, as well as text verification to ensure clarity and grammatical correctness and generating the pipeline overview image. The use of Generative AI was instrumental in accelerating the development process, providing insights, and enhancing the overall quality of the work.

4.2 GENERAL VIEW OF THE PROPOSED APPROACH

The proposed approach for fine-tuning a large language model to extract structured information from legal documents is illustrated in Figure 2. The pipeline begins with the input of raw documents from legal documents, which are then annotated through a hybrid approach. Later the annotated dataset is used to fine-tune a 4-billion-parameter model (Gemma 3) using quantization and Low-Rank Adaptation (LoRA). The two proposed fine-tuning approaches are evaluated and compared, with the best performing model being selected as the final model (LegalExtract470). The final model is then evaluated on a held-out test set to assess its performance in extracting structured information from legal documents.

Figure 2 – Pipeline Overview



4.3 TASK FORMULATION

The extraction task is a sequence to sequence task that consists of a structured sequence to sequence problem. Given a raw text excerpt from a legal document, specifically case files of administrative impropriety, the model is required to generate a strictly formatted JSON object.

The expected JSON output consists of three primary root keys: "DADOS PROCESSO", "DADOS DOS REQUERIDOS", and "DADOS LICITAÇÃO". These correspond to information regarding the civil case file, the defendants, and the public bidding process, respectively. Each of these fields maps to structured data as defined in Tables 2 and 3. With the exception of the defendants' information ("DADOS DOS REQUERIDOS"), which expects a list of JSON objects (each detailing the ID, name, and role of an individual defendant), the extracted values are standard key-value pairs. If the text excerpt lacks the necessary information to populate a specific fields, the model is instructed to output the string "Null".

To illustrate the lexical complexity and the semantic abstraction required by this task, Tables 1 and 2 presents a concrete, anonymized example of a raw text input from a case file and its corresponding expected JSON output. This "before and after" comparison highlights why a standard exact-match token loss is insufficient: the model must map varied textual formulations (such as legal jargon for overbilling or financial damages) into a rigid, unified schema. This semantic gap is what fundamentally motivates the implementation of the Custom Loss Function detailed in later sections.

Table 1 – Example of Input Text

[alvo] III - FATOS. Segundo consta da documentação que acompanha a presente petição inicial, notadamente a cópia do Inquérito Civil n. 06.2017.00005825-9, <name 1>, na condição de Secretário Municipal de Saúde do Município de Içara, acusou o recebimento dos materiais referentes às notas fiscais de n. 30989, 30991, 30993, 30994, 31757, 31758, 31759, 31760, 31761, 31762, 31763, 31764 em favor da empresa <name 2> e das notas fiscais n. 4694, 4741, 4820 em favor da requerida <name 3>, sem que os produtos tenham sido efetivamente entregues, o que gerou dano ao erário municipal e enriquecimento ilícito das empresas. O município de Içara, por meio dos pregões presenciais n. <id 1>, n. <id 2> e n. <id 3>, realizou a aquisição de materiais de enfermagem e odontológicos. O pregão n. <id 1> foi aberto no dia <date 5> e teve como objeto a contratação de empresa especializada do ramo pertinente para a aquisição de materiais de consumo e permanente de enfermagem para as unidades de saúde, com valor estimado de R\$ <value 1>. A abertura dos envelopes ocorreu no dia <date 8>, sendo licitados 166 itens. Onze empresas participaram da licitação, resultando a contratação, dentre outras, das empresas ora requeridas. <name 3> foi contratada no valor de R\$ <value 2> por meio do contrato n. <contract 3> e <name 2> foi contratada no valor de R\$ <value 3> por meio do contrato n. 36/2016. Os contratos foram assinados em <date 6> e tinham vigência até <date 7>. O pregão presencial n. <id 2> foi aberto em 6/4/2016 e teve como objeto a contratação de empresa especializada do ramo pertinente para a aquisição de materiais de consumo odontológico para unidades de saúde e para o centro de especialidade odontológicas. Neste processo foram licitados 369 itens, sendo o critério de julgamento o menor preço por item. A abertura das propostas foi em <date 4> e a adjudicação e homologação foi em 9/5/2016. Três empresas participaram do certamente e o resultado final ensejou a contratação, dentre outras, das empresas Rés. <name 2> foi contratada no valor de R\$ <value 4> por meio do contrato n. <contract 1> e <name 3> restou contratada no valor de R\$ <value 5> por meio do contrato n. <contract 2>. Ambos os contratos tinham vigência até <date 7>. O pregão presencial n. <id 3>, por sua vez, foi aberto em 16/<id 1> e teve como empresa vencedora, dentre outras, a requerida <name 3>. A empresa foi contratada no valor de R\$ <value 6> por meio do contrato n. <contract 4>. O requerido <name 1> ocupou o cargo de <role 1>. No período em que <name 1> esteve no comando da Secretaria de Saúde do Município de Içara, foram noticiadas ausência de materiais nas unidades de saúde da municipalidade. Após investigações, verificou-se que parte dos materiais acusados como recebidos pelo então secretário não foram de fato entregues à municipalidade.

Source: Compiled by the author (2026).

Table 2 – JSON schema definition for case/case file extraction.

Key	Value Definition
INQUÉRITO ASSOCIADO	ID of the civil case
SIG (RELACIONADO AO INQUÉRITO)	SIG ID tied to the civil case
COMARCA DE AJUIZAMENTO DA AÇÃO	County where the case was listed
DATA DE PUBLICAÇÃO DO PROCESSO	Date when the case was filed
VALOR DA CAUSA	Financial value of the claim
TESTEMUNHAS	List of witnesses
TIPO DE IMPROBIDADE	Type(s) of administrative impropriety
PROMOTOR	Prosecutor(s) responsible for the document

4.4 DOCUMENT STRUCTURE

Because the source documents originate from different counties and follow varying procedural norms—and are not exclusively civil cases—their structural formatting

Table 2 – Expected JSON Output for the Input Text

```
[saida]
{
  "DADOS PROCESSO": {
    "INQUÉRITO ASSOCIADO": "06.2017.00005825-9",
    "SIG (RELACIONADO AO INQUÉRITO)": "Null",
    "COMARCA DE AJUIZAMENTO DA AÇÃO": "Null",
    "PROMOTOR": "Null",
    "DATA DE PUBLICAÇÃO DO PROCESSO": "Null",
    "VALOR DA CAUSA": "Null",
    "TIPO DE IMPROBIDADE": "Null",
    "TESTEMUNHAS": "Null"
  },
  "DADOS DOS REQUERIDOS": [
    {
      "NOME": "<name 1>",
      "FUNÇÃO": "<role 1>"
    },
    {
      "NOME": "<name 2>",
      "FUNÇÃO": "empresa favorecida"
    },
    {
      "NOME": "<name 3>",
      "FUNÇÃO": "empresa favorecida"
    }
  ],
  "DADOS LICITAÇÃO": {
    "NÚMERO PROCESSO LICITATÓRIO": "Null",
    "NÚMERO EDITAL(PROCESSO LICITATÓRIO)": [
      "<id 1>",
      "<id 2>",
      "<id 3>"
    ],
    "DATA DE PUBLICAÇÃO DO EDITAL": [
      "<date 1>",
      "<date 2>",
      "<date 3>"
    ],
    "VALOR DO SERVIÇO": "<value 1>",
    "TIPO DE PRODUTO-SERVIÇO": "materiais de consumo e
      permanente de enfermagem para as unidades de saúde",
    "MODALIDADE": [
      "Pregão Presencial",
      "Pregão Presencial",
      "Pregão Presencial"
    ]
  }
}
```

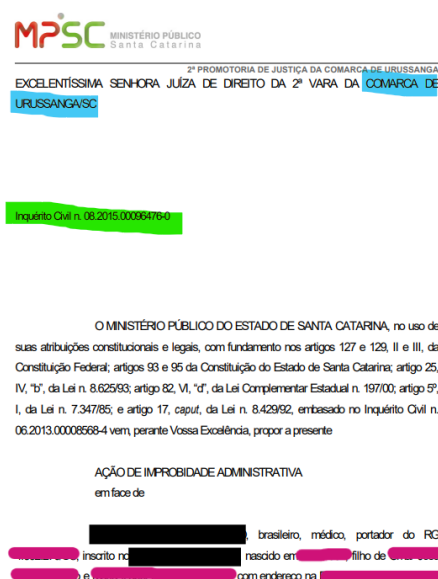
Source: Compiled by the author (2026).

Table 3 – JSON schema definition for public bidding extraction.

Key	Value Definition
NÚMERO PROCESSO LICITATÓRIO	ID of the overall bidding process
NÚMERO EDITAL (PROCESSO LICITATÓRIO)	ID of the specific bidding notice
DATA DE PUBLICAÇÃO DO EDITAL	Registration date of the bidding
VALOR DO SERVIÇO	Cost associated with the contract
SUPERFATURAMENTO	Estimated cost of overbilling
TIPO DE PRODUTO-SERVIÇO	Goods and services contracted

varies significantly. Nevertheless, there are a few structural certainties present across the corpus. First, it is highly probable that every page of the document contains a header indicating the county of origin (highlighted in blue in Figure 3); this is guaranteed on the first page. Second, the document typically features an identification block, such as the civil case ID or the SIG (highlighted in green). Following this, there is a list of defendants containing their names and descriptions (censored in black), which may optionally include their personal identification numbers and professions.

Figure 3 – Beginning of a Document



As seen in Figure 4, the end of each document contains <claims>, that will occasionally reference the type of administrative impropriety committed by the demanded (highlighted in pink), the names and roles of the defendants (censored in black), and the value of the claim (highlighted in blue). Followed by the county where the document was produced (orange), date (green) and the name of the prosecutor responsible for said document (yellow).

Figure 4 – End of a Document

seja condenado, por infringência ao artigo 9º, caput e inciso I da Lei n. 8.429/92, nas sanções previstas no artigo 12, I, daquela mesma Lei.

g) subsidiariamente, caso improcedente o pedido anterior, postula-se a procedência do pedido para a condenação de por infringência ao artigo 11, caput e inciso I da Lei n. 8.429/92, nas sanções previstas no artigo 12, III, daquela mesma Lei;

h) a condenação do réu ao pagamento das despesas processuais e demais verbas de sucumbência.

Pleiteia também a este juízo que comunique, por ofício, o Conselho Regional de Medicina sobre o ajuizamento da presente ação, para que administrativamente tome as providências cabíveis.

Dá-se à causa o valor de R\$ 1.300,00 (mil e trezentos reais).

Urussanga, 13 de abril de 2015.

Claudine Vidal de Negreiros da Silva

Promotora de Justiça

4.5 DATASET ANNOTATION

For this work, the raw data used to produce the dataset was provided by Projeto Ceos, a University of Santa Catarina’s Department of Informatics and Statistics initiative that is a part of the MPSC (Ministério Público de Santa Catarina). Through the partnership the lab produces multiple database, data mining, machine learning and AI projects for the MPSC. The dataset was produced by manually annotating a subset of 100 legal documents, which were randomly selected from a larger pool of over 5000 documents provided by Projeto Ceos. The annotation process involved extracting relevant information from the legal texts and structuring it into the predefined JSON format, as described in the previous section. From those, the first 50 documents were randomly selected and hand annotated one by one (seen in Figure 5).

Figure 5 – Fragment of Annotation

Entrada	Saída	Parâmetro
None	None	Inquérito associado
None	None	SIG (relacionado ao inquérito)
EXCELENTÍSSIMO SENHOR DOUTOR JUIZ DE DIREITO DA 1ª VARA CÍVEL DA COMARCA DE SÃO BENTO DO SUL, O MINISTÉRIO PÚBLICO DO ESTADO DE SANTA CATARINA, por seu representante, com base nos artigos 127 e 129, inciso III, da Constituição Federal de 1988, artigo 25, inciso IV, alínea "b", da Lei n. 8.525/93, artigo 82, inciso VI, alínea "d", da Lei Complementar...	SÃO BENTO DO SUL	COMARCA DE AJUIZAMENTO DA AÇÃO
1. OBJETIVO DA AÇÃO Esta ação objetiva a instrução e aplicação de sanções decorrentes de ato improbo perpetrado, no exercício de suas funções, pelos Policiais Militares	São Bento do Sul	MUNICÍPIO DE IRREGULARIDADE
previous=2	None	NOMES E CPFs PESSOAS ENVOLVIDAS
previous=2	None	NOMES E CPFs PESSOAS ENVOLVIDAS
previous=2	None	NOMES E CPFs PESSOAS ENVOLVIDAS
previous=2	Cabo PM	NOMES E FUNÇÕES PESSOAS ENVOLVIDAS
previous=2	Soldado PM	NOMES E FUNÇÕES PESSOAS ENVOLVIDAS
previous=2	Soldado PM	NOMES E FUNÇÕES PESSOAS ENVOLVIDAS
Acerca da violação desses princípios, leciona Celso Antonio Bandeira de Mello: "Violar um princípio é muito mais grave que transgredir uma norma qualquer. A desatenção ao princípio implica ofensa não apenas a um específico mandamento obrigatório mas a todo sistema de comandos. É a mais grave forma de legalidade...	art. 11 da Lei n. 8.429/92	TIPO DE IMPROBIDADE

The next 50 were pre-annotated by Llama 3.3 (70 Billion parameters) on Ceos’ ollama server. Looking to achieve good performance in this pre-annotation stage the model was fed a detailed prompt reinforcing the structure of the output and applying 10-

shot prompting. The examples used in prompting came from the previously annotated examples. From this pre-annotated state, the author manually verified and corrected every instance. Although the author is not a specialist in this specific domain, careful cross-referencing and individual review were applied to integrate a high-quality dataset.

4.6 SETUP

Training was performed on the Jupyter Notebook servers at the Federal University of Santa Catarina (UFSC), utilizing a single NVIDIA H100 GPU. The implementation was built using the Hugging Face libraries, specifically employing a specialization `SFTTrainer` class that integrates the custom loss function defined in Equation 4.1.

4.7 QUANTIZATION AND MODEL INITIALIZATION

To enable the fine-tuning of Gemma 3, a Large Language Model comprising 4 billion parameters, within the constraints of available computational resources, weight quantization was employed. This specific foundation model was selected due to its strong baseline performance on complex information extraction tasks.

The model was initialized utilizing the Hugging Face transformers and bitsandbytes libraries. Specifically, the base model weights were quantized to 4-bit NormalFloat (NF4) precision. To further optimize memory usage without severely degrading representational capacity, double quantization was enabled, and the computation data type during the forward pass was set to 16-bit floating-point (FP16). This configuration drastically reduces the GPU VRAM footprint, allowing the 4 billion parameters model to be loaded and trained efficiently while retaining its precision during matrix multiplications.

4.8 LOW-RANK ADAPTATION (LORA) CONFIGURATION

Parameter-Efficient Fine-Tuning was implemented via Low-Rank Adaptation (LoRA) using the Hugging Face `peft` library. The model was configured for a standard Causal Language Modeling (Causal LM) task.

To strike an optimal balance between the number of trainable parameters and model plasticity, the low-rank adapter matrices were initialized with a rank of $R = 16$ and an alpha scaling factor of $\alpha = 32$. A dropout rate of 10% was applied to the adapter layers to mitigate overfitting on the relatively small training dataset.

Crucially, rather than restricting the adapters to only the attention query and value projections, LoRA adapters were injected comprehensively across the Transformer architecture. The targeted modules included all self-attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`) as well as the Multi-Layer Perceptron (MLP) components (`up_proj`,

down_proj, and gate_proj). Targeting this comprehensive set of modules maximizes the model's capacity to adapt to the highly specific structural JSON formats and semantic patterns inherent to the legal documents.

4.9 SUPERVISED FINE TUNING DESCRIPTION

Models were trained on subsets of sizes 100, 200, 300, 400, and finally 470 examples. The validation dataset was composed of 20% of training examples. The examples in the datasets are randomly selected from the set of all available examples. Important hyperparameters are set as described in table 4. Because of hardware limitations, the batch size was set to 4, with 4 gradient accumulation steps, resulting in an effective batch size of 16. The learning was adjusted to 3×10^{-4} empirically while testing values within the range of $[1 \times 10^{-5}, 1 \times 10^{-3}]$. The learning rate scheduler was set to Cosine, with a maximum sequence length of 2048 tokens (also due to hardware limitations). The model was trained for a maximum of 8 epochs, which was determined to be sufficient for convergence based on preliminary experiments as the model tended to overfit quickly. Due to the small size of the dataset, early stopping was applied with a patience of 3 epochs, halting training if the validation loss did not improve for 3 consecutive epochs. The optimizer used was Paged 8-bit AdamW, which is well-suited for large language models and helps in stabilizing training dynamics.

Table 4 – Parameters for Supervised Fine Tuning

Parameter	Value
Batch size	4
Gradient Accumulation Steps	4
Learning Rate	3×10^{-4}
Learning Rate Scheduler	Cosine
Maximum Sequence Length	2048
Number of Training Epochs	8
Optimizer	Paged 8-bit AdamW
Early stopping patience	3

4.9.1 Custom Loss Function

When extracting information from text using LLMs, standard next-token prediction metrics (Cross-Entropy) heavily penalize outputs that are lexically different but semantically equivalent to the ground-truth label (e.g., using synonyms or paraphrasing a legal clause). Furthermore, standard Cross-Entropy does not explicitly enforce the hierarchical structure of JSON formats. To properly reward semantic equivalence and structural fidelity during the initial supervised fine-tuning phase, the model incorporates a custom multi-objective loss function, defined in Equation 4.1.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{tree}}(t)\mathcal{L}_{\text{tree}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{struct}}\mathcal{L}_{\text{struct}} \quad (4.1)$$

The total loss $\mathcal{L}_{\text{total}}$ is a weighted sum of four distinct components:

1. **Cross-Entropy Loss ($\mathcal{L}_{\text{ce}}$):** The standard autoregressive token-level negative log-likelihood loss.
2. **Semantic Loss ($\mathcal{L}_{\text{sem}}$):** To maintain differentiability without breaking the computational graph, the model computes an expected continuous embedding for each predicted token by weighting the vocabulary embeddings by their Top- K softmax probabilities (where $K = 8$). These expected embeddings and the ground-truth token embeddings are mean-pooled across the sequence length. The semantic loss is defined as the cosine distance ($1 - \text{sim}_{\text{cos}}$) between the pooled predicted representations and the ground-truth representations, penalizing semantic divergence.
3. **Tree Alignment Loss ($\mathcal{L}_{\text{tree}}$):** To encourage the generated sequence to match the hierarchical length and structure of the target JSON, an Optimal Transport (OT) approach is utilized. The sequence embeddings are divided into continuous chunks and pooled. A pairwise cost matrix is calculated using cosine distance between the predicted chunks and ground-truth chunks. The loss is then computed using the Sinkhorn-Knopp algorithm, which efficiently approximates the Earth Mover’s Distance between the two sequences. Because this alignment constraint can destabilize early training, its weight $\lambda_{\text{tree}}(t)$ is dynamically warmed up over the first training steps based on the global step (each time the loss is computed for a batch) t in a manner $\frac{t}{\text{warmup_steps}}$.
4. **Structural Loss ($\mathcal{L}_{\text{struct}}$):** To explicitly enforce JSON schema generation, this component boosts the probability of fundamental JSON syntax characters. Let $\mathcal{S} = \{\{, \}, [,], :, ', '\}$ be the set of structural tokens. The loss penalizes the model if the predicted probability distributions assign low likelihoods to these critical tokens across the sequence:

$$\mathcal{L}_{\text{struct}} = \frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \mathbb{E} [1 - P(v)] \quad (4.2)$$

Based on empirical tuning, the static scalar weights balancing these objectives were set to $\lambda_{\text{ce}} = 0.65$, $\lambda_{\text{sem}} = 0.55$, and $\lambda_{\text{struct}} = 0.30$, with the optimal transport alignment weight λ_{tree} warming up to a maximum of 0.50.

4.10 REINFORCEMENT LEARNING (TWO-STAGE APPROACH)

As an alternative to purely supervised training, a two-stage training paradigm was evaluated. Instead of inducing strict JSON schema adherence and extraction correctness solely through Supervised Fine-Tuning (SFT), this second approach utilizes an initial SFT phase over the full sequence using standard Cross-Entropy loss. This first stage equips the model with general extraction capabilities and a baseline understanding of JSON formatting. Subsequently, Reinforcement Learning (RL) is applied in a second stage to rigorously align the model's outputs with the required structural and semantic constraints.

4.10.1 DAPO Configuration and Training Dynamics

To conduct the reinforcement learning phase of the second training approach, the Dynamic Sampling Policy Optimization (DAPO) algorithm was adopted to update the model's policy. Leveraging the group-relative advantage framework detailed previously, the training environment was configured to generate a group of $G = 4$ distinct completions for each input prompt. The programmatic reward function evaluates each completion, and the algorithm calculates the training advantage by normalizing these scores against the group's mean and standard deviation.

To maintain linguistic coherence and prevent "reward hacking"—where the model artificially maximizes embedding similarity at the expense of generating valid text—a Kullback-Leibler (KL) divergence penalty (β) was enforced. This penalty regulates the degree of freedom the optimization process has to differentiate the active policy from the baseline distribution acquired during the SFT stage. In order to evaluate the optimal balance between structural exploration and prior knowledge retention, multiple values for the KL divergence penalty were tested, with $\beta = 8.5$ serving as the primary configuration. Finally, the model was fine-tuned using a learning rate of 1×10^{-5} and 16 gradient accumulation steps. This configuration simulated a larger effective batch size, promoting stable optimization over the long token sequences inherent to complex JSON extraction.

4.10.2 Programmatic Reward Shaping

Because data extraction into a specific schema presents objective success criteria, human preference labeling is not required. Instead, a programmatic reward function was engineered to evaluate each generated output during the RL phase. The final reward evaluates two primary criteria: formatting validity and semantic correctness.

First, the **Format Reward** (R_{format}) functions as a strict syntactic gatekeeper. It is a binary function that returns 1.0 if the generated text can be successfully parsed into a

valid JSON object, and 0.0 otherwise.

Second, to prevent the model from being penalized for valid synonymous extractions or minor lexical deviations, a **Semantic Reward** (R_{semantic}) is applied. The generated JSON prediction (y_{pred}) and the ground truth (y_{gt}) are flattened into sorted string representations. These strings are then mapped into a dense vector space using a lightweight Sentence Transformer (all-MiniLM-L6-v2), denoted as the embedding function $E(\cdot)$. The cosine similarity between these embeddings is calculated and linearly mapped from the $[-1, 1]$ range to a $[0, 1]$ range, as defined in Equation 4.3.

$$R_{\text{semantic}}(y_{\text{pred}}, y_{\text{gt}}) = \frac{\text{sim}_{\cos}(E(y_{\text{pred}}), E(y_{\text{gt}})) + 1}{2} \quad (4.3)$$

The final scalar reward for each generation, R_{total} , is computed as a weighted linear combination of these components (Equation 4.4), where w_1 and w_2 act as scalar weights controlling the importance of syntax versus semantics. The resulting reward is explicitly clamped to the $[0, 1]$ interval to guarantee numerical stability during the policy updates.

$$R_{\text{total}} = \text{clip}(w_1 R_{\text{format}} + w_2 R_{\text{semantic}}, 0, 1) \quad (4.4)$$

4.11 EXPERIMENT DESIGN

The primary objective of the proposed experiments is to evaluate the performance of the final fine-tuned model on the structured extraction task. Specifically, the experiments are designed to address the following goals:

- **Goal 1: Baseline Comparison.** Compare the extraction accuracy and structural consistency of LegalExtract470 against the original zero-shot and few-shot Gemma 3 baselines.
- **Goal 2: Generalization Assessment.** Evaluate the models' performance on unseen data by testing them on a held-out subset of one hundred examples from the annotated dataset.
- **Goal 3: Multi-Dimensional Evaluation.** Score and analyze the models' generations across different dimensions of quality, ranging from structural integrity to lexical precision.
- **Goal 4: Robustness Assessment** Verify the robustness of the models across different data types (e.g., highly structured dates vs. unstructured witness lists) through a granular token-level analysis."

To achieve these goals, the models' generations were scored across six main metrics, each implemented to capture a distinct aspect of performance:

- **Tree Edit Distance (TED):** Evaluates the structural and hierarchical integrity of the generated JSON. Both the predicted and ground-truth JSON objects are first converted into tree structures. The All-Path Tree Edit Distance (APTED) algorithm is then used to compute the minimum number of node operations (insertions, deletions, and substitutions) required to transform the prediction tree into the reference tree. In our case, since the order of the keys does not matter, both the predicted json and the ground truth keys are sorted. Finally, this raw distance is then normalized by the size of the ground-truth object to produce a structural similarity score bounded between 0 and 1, where 1 signifies a perfect structural match.
- **F1-Score (F_1):** Computes the harmonic mean of precision and recall at the key-value field level. To enable this calculation, both the predicted and ground-truth JSON objects are first flattened into a set of unique (path, value) string pairs, where the path component does not refer to ordering. The F1-score is then calculated at the micro level, where true positives, false positives, and false negatives are determined by exact matches within these sets. This provides a single score reflecting the overall accuracy of the extracted fields.
- **Validity Rate (v_r):** Measures a model's capability to consistently output valid JSON. It represents the proportion of generations that contain a well-formed JSON string that can be successfully parsed by a standard decoder, after extracting the content between the first { and last } characters. It is defined simply as the ratio of valid outputs to total outputs:

$$v_r = \frac{\text{valid}}{\text{total}}$$

- **Semantic Similarity (SSim):** Quantifies the semantic equivalence of the extracted values, independent of the JSON structure, to account for valid paraphrasing. The values from the predicted and ground-truth JSON objects are extracted and encoded into embedding matrices using a Sentence Transformer model. A cosine similarity matrix is then computed between these two sets of embeddings. The final score is calculated by finding the maximum similarity score for each predicted value against all ground-truth values, and then averaging these maximum scores. This metric evaluates the model's capability to generate semantically correct values, even when the key-value pair is misaligned.
- **Consistency (Consis):** Quantifies the reliability and stability of the model's structured extractions by measuring the structural and textual alignment across multiple

independent generation cycles. Raw outputs from repeated inference runs are first cleaned and parsed into structured JSON objects, filtering out any malformed objects. Assuming at least two valid structures are successfully generated, the metric executes a combinatorial pairwise comparison that evaluates every unique pair of JSON outputs exactly once. For each pair, a dedicated F1-score function measures the precise overlap of extracted keys and values, balancing token precision against recall. The final consistency score is derived by computing the arithmetic mean of these pairwise scores, yielding a value between 0.0 and 1.0 that isolates the model's deterministic stability and robustness against prompt ambiguity, independent of its absolute alignment with ground-truth data.

- **Token Accuracy:** It is the ratio of perfectly matched tokens between the generated and ground-truth JSON outputs. This metric captures the model's ability to produce exact lexical matches for each token in the output, providing a fine-grained measure of correctness at the token level.

4.12 PROMPTING

In addition to the six metrics, we prepared two different prompts (all formulated in Portuguese to match the language of the target legal documents) in order to fully assess the capabilities of the baseline models. The first prompt describes the task and provides the definition of the expected JSON schema. The second prompt (Table 3) improves on this version by making use of in-context learning, containing three examples of input-output pairs extracted from the training set.

4.13 VALIDATION

Validation of the tuned models was performed using the same metrics and prompts as described in the experiments section. The validation set, consisting of 20% of the training examples, was used to monitor the performance of the models during training and to select the best checkpoint based on the highest performing model on the main metrics (F1-score, Semantic Similarity and Validity Rate). This approach ensures that the selected model not only performs well on the training data but also generalizes effectively to unseen examples, as reflected in the validation metrics. This validation process was used to select hyperparameters and to select the final model checkpoint for evaluation on the held-out test set.

Table 3 – Target JSON schema defined for the administrative impropriety extraction task.

System Prompt Instructions
Extraia dados do texto marcado por [alvo] e estuture a saída na forma de um json após [saída]. Os principais campos do json são: DADOS PROCESSO, DADOS DOS REQUERIDOS e DADOS LICITAÇÃO. Os dados adicionados a DADOS PROCESSO são um json menor com dados do documento, assim como DADOS LICITAÇÃO. Desse modo, o json de saída deve seguir a estrutura: <pre> { "DADOS PROCESSO": { "INQUÉRITO ASSOCIADO": "str", "SIG (RELACIONADO AO INQUÉRITO)": "str", "COMARCA DE AJUIZAMENTO DA AÇÃO": "str", "PROMOTOR": ["str", "list[str]"], "DATA DE PUBLICAÇÃO DO PROCESSO": "str", "VALOR DA CAUSA": "float", "TIPO DE IMPROBIDADE": ["str", "list[str]"], "TESTEMUNHAS": ["str", "list[str]"] }, "DADOS DOS REQUERIDOS": ["str", "list[objeto_requerido]"], "DADOS LICITAÇÃO": { "NÚMERO PROCESSO LICITATÓRIO": ["str", "list[str]"], "NÚMERO EDITAL(PROCESSO LICITATÓRIO)": ["str", "list[str]"], "DATA DE PUBLICAÇÃO DO EDITAL": ["str", "list[str]"], "VALOR DO SERVIÇO": ["str", "list[str]"], "TIPO DE PRODUTO-SERVIÇO": ["str", "list[str]"], "MODALIDADE": ["str", "list[str]"] } } objeto_requerido = { "NOME": "str", "FUNÇÃO": "str", "CPF": "str" } </pre> O campo "TIPO DE IMPROBIDADE" deve ser preenchido com o artigo da lei de improbidade que descreve o tipo de infração (art. 11 da Lei n. 8.429/92 para atentado aos princípios da administração pública, art. 10 da Lei n. 8.429/92 para dano ao erário e art. 9 da Lei n. 8.429/92 para enriquecimento ilícito). Caso a informação necessária para preencher algum dos campos não esteja presente no texto ele deve ser preenchido com a string "Null".

Source: Compiled by the author (2026).

5 RESULTS

The results of the experiments conducted in this study are presented in this chapter. The performance of the fine-tuned models is evaluated using a set of metrics that assess both the semantic accuracy and structural consistency of the extracted JSON data. The evaluation is performed on a held-out test set, which was not used during the training or validation phases.

5.1 COMPARISON OF THE TWO TUNING APPROACHES

Table 5 – Tuned models’ performance comparison on the structured legal data extraction task.

Model	TED	F1-Score	SSim	Valid Rate	Consis
Pure SFT (100 examples)	0.7416	0.6246	0.6842	0.7500	0.4573
Pure SFT (200 examples)	0.7892	0.6486	0.7296	0.8000	0.6187
Pure SFT (300 examples)	0.7897	0.6346	0.7318	0.8000	0.5773
Pure SFT (400 examples)	0.8763	0.7362	0.8174	0.8875	0.6509
Pure SFT (470 examples)	0.9170	0.8280	0.9114	0.9255	0.7222
SFT + RL	0.9287	0.8095	0.8856	0.9362	0.6920

As seen in Table 5, the pure SFT approach with 470 examples achieved the best performance across multiple metrics on the validation set. The SFT + RL approach did not outperform the pure SFT method in this validation phase, indicating that the additional complexity of reinforcement learning in this restricted training set may not have provided a significant benefit for this specific task and dataset size. Based on these results, the pure SFT model trained with 470 examples was selected as the final model checkpoint for evaluation on the held-out test set.

5.2 TEST SET EVALUATION

This section presents the performance evaluation of the baseline models and the fine-tuned model, LegalExtract470, on the held-out test set. The quantitative results across the six evaluation metrics are summarized in Table 6, while Figure 9 provides a visualization of the multi-metric trade-offs using a radar chart.

First, we observe the impact of In-Context Learning on the baseline foundation models. Comparing the zero-shot baseline Gemma3:4B (raw) to its three-shot counterpart, Gemma3:4B (3-shot), reveals a substantial performance improvement across nearly every metric, with the most notable increases occurring in F1-score and token accuracy. Nonetheless, the smaller 3B parameter model (Qwen2.5:3B) struggled with extraction

Table 6 – Global performance comparison of baseline models and the proposed LegalExtract470 model on the structured legal data extraction task.

Model	TED	F1-Score	SSim	Valid Rate	Consis	Token Acc.
Qwen2.5:3B (3-shot)	0.9784	0.6807	0.8203	1.0000	0.7243	0.0097
Gemma3:27B (3-shot)	0.9859	0.7947	0.9229	1.0000	0.9821	0.4969
Gemma3:4B (3-shot)	0.9733	0.7488	0.8909	0.9900	0.9353	0.3862
Gemma3:4B (raw)	0.9641	0.4548	0.8273	1.0000	0.8654	0.1559
LegalExtract470	0.9516	0.8250	0.9021	0.9600	0.7254	0.4173

Figure 6 – Semantic Accuracy "DADOS LICITAÇÃO"

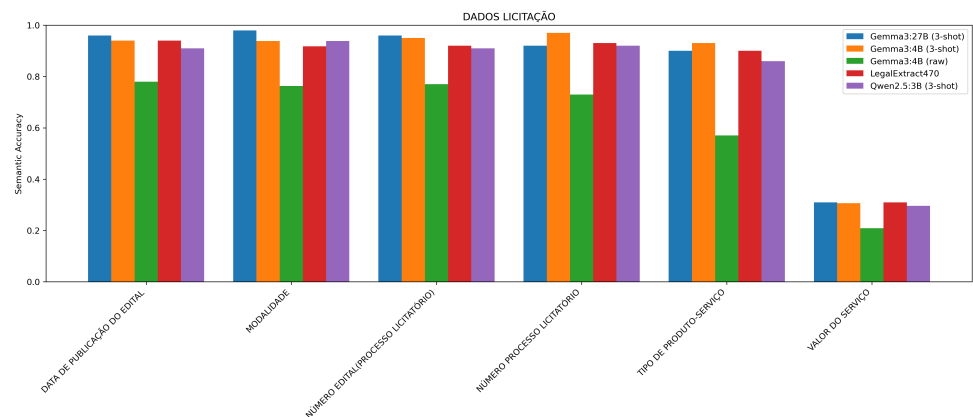


Figure 7 – Semantic Accuracy "DADOS PROCESSO"

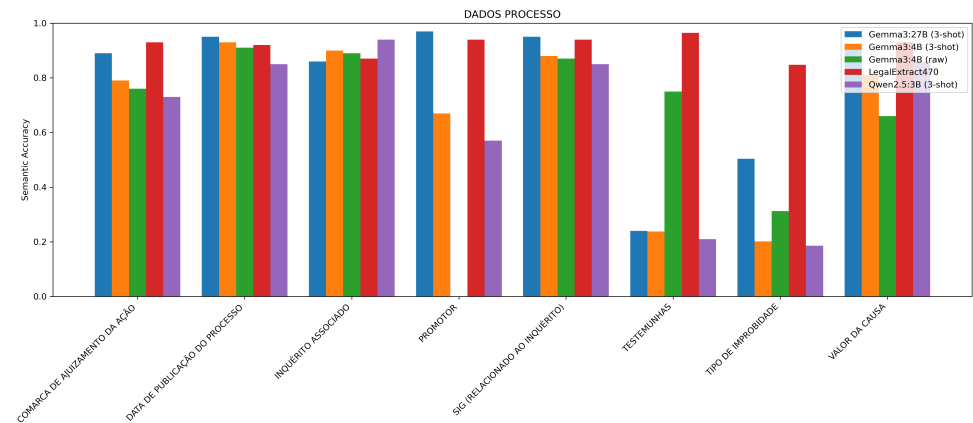


Figure 8 – Semantic Accuracy "DADOS DOS REQUERIDOS"

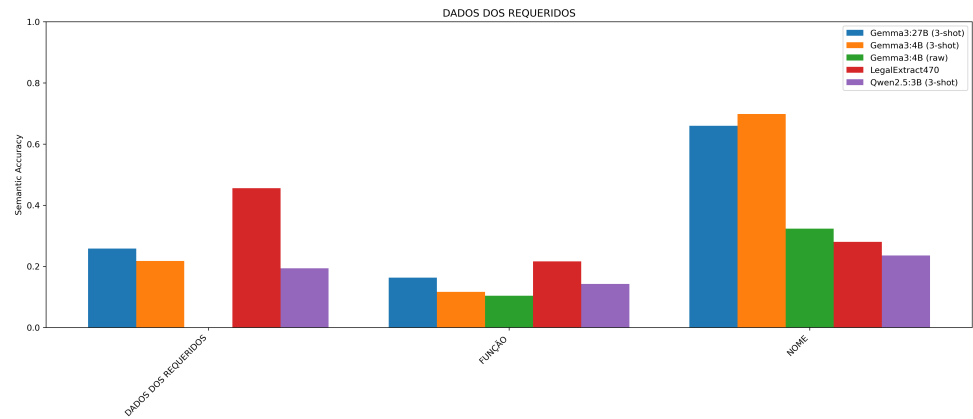
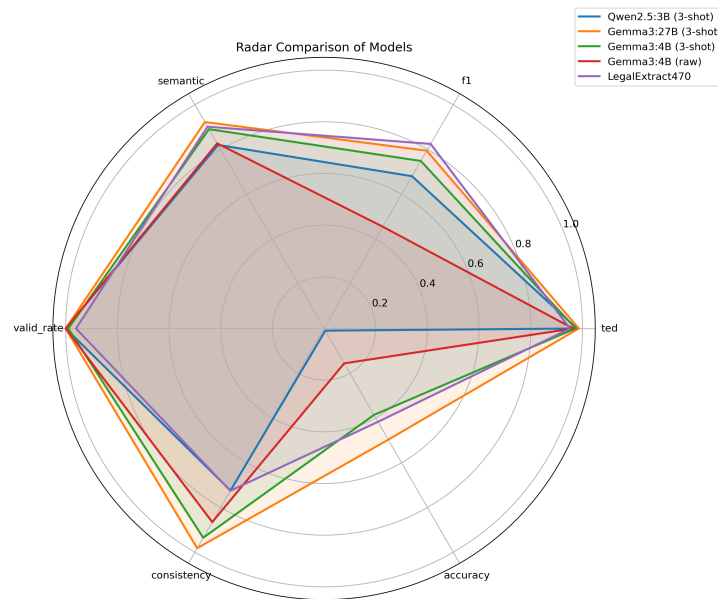


Figure 9 – Radar View of Global Metrics



accuracy, producing a token match score of only 0.0097, despite its perfect valid JSON rate.

Second, the proposed fine-tuned model, LegalExtract470, demonstrates strong performance in factual retrieval. It achieves the highest F1-score (0.8250) among all evaluated models, including the significantly larger Gemma3:27B (3-shot) baseline (0.7947). Our model also produces a competitive semantic similarity score of 0.9021 and an token match accuracy of 0.4173. This indicates that across all valid JSON structures generated during testing, over 41% of the individual tokens perfectly matched their corresponding target tokens. Consequently, LegalExtract470 outperforming other baseline models of comparable scale at a lower computational cost.

However, as illustrated by the radar graph in Figure 9, parameter-efficient fine-tuning introduced a notable trade-off between semantic extraction capability and structural consistency. Because the consistency evaluation employs a higher sampling temperature, the fine-tuned model becomes susceptible to autoregressive text degeneration. When a low-probability or suboptimal token is sampled early in the generation sequence, the model enters a self-reinforcing repetition loop, effectively losing its contextual footing and continuously generating degenerate formatting. Consequently, while LegalExtract470 excels at identifying and correctly retrieving specific legal key-value pairs, it exhibits a slight regression in both structural consistency and JSON validity (dropping to 96%) when compared to the three-shot baselines.

5.2.1 Per-Field Granular Analysis

To better understand the practical performance of the models within the Brazilian administrative impropriety domain, a granular analysis of extraction accuracy at the token level was conducted. Figure 6, Figure 7, and Figure 8 present semantic accuracy (an answer is deemed correct in this case if its cosine similarity to the expected value for the field is over 0.80) for each field in the DADOS LICITAÇÃO, DADOS PROCESSO, and DADOS DOS REQUERIDOS sections, respectively. Each graph represents the semantic accuracy for a specific field across all evaluated models, with brighter colors indicating higher accuracy.

The graphs reveals several key architectural behaviors:

1. **Systematically mastered fields:** Highly structured metadata, such as dates (DATA DE PUBLICAÇÃO DO PROCESSO, DATA DE PUBLICAÇÃO DO EDITAL), case IDs (INQUÉRITO ASSOCIADO, SIG), and bidding modalities (MODALIDADE), are reliably extracted across almost all models, as represented by the bright yellow bands. These fields generally follow strict formatting rules for legal headers, making them easily identifiable.
2. **Complex, unstructured fields:** A key finding concerns the performance on the DADOS PROCESSO.TESTEMUNHAS (witnesses) and DADOS PROCESSO.TIPO DE IMPROBIDADE (type of impropriety) fields. Although zero-shot and few-shot baselines largely fail to extract witness lists from legal prose and differentiate them from DADOS DOS REQUERIDOS entries, LegalExtract470 achieves an exceptionally high token accuracy (represented by the bright yellow highlight in its column). This indicates that fine-tuning successfully enabled the model to learn a more nuanced approach to this extraction task.
3. **Inherent dataset bottlenecks:** Finally, fields like DADOS PROCESSO.MUNICÍPIO DE IRREGULARIDADE or the more complex nested structure in DADOS DOS REQUERIDOS represent a clear bottleneck in the performance of all models. This limitation may stem from the inherent nature of these fields. For instance, DADOS PROCESSO.MUNICÍPIO DE IRREGULARIDADE is easily confused with COMARCA DE AJUIZAMENTO DA AÇÃO (county of filing) without a careful analysis of the entire document's context. Similarly, DADOS DOS REQUERIDOS.FUNÇÃO (respondent's role) is a complex field because it may not be present in the paragraph where the respondent is first introduced. Furthermore, its correct interpretation requires a broader understanding of the illegalities described throughout the entire document, which is never available to the model as we are constrained by the sequence length during training. Additionally, this outcome may stem from a lack of sufficient training examples for specific fields, particularly given that legal documents can be highly heterogeneous.

Overall, the experimental evaluation demonstrates that the proposed parameter-efficient fine-tuning pipeline successfully adapts a compact language model to the task of structured legal information extraction. The results indicate that supervised fine-tuning alone was sufficient to produce the strongest model under the available training conditions, while the detailed analyses highlighted both the strengths of the proposed approach and the remaining limitations associated with long-context reasoning, structural consistency, and challenging semantic fields. These findings provide a comprehensive answer to the research questions posed in Chapter 1 and establish the foundation for the broader discussion presented in the following chapter, where the contributions, limitations, and future research directions of this work are examined.

6 CONCLUSION AND FUTURE WORK

This work aimed to address the challenging and labor-intensive task of structured data extraction from heterogeneous Brazilian legal documents. By proposing a parameter-efficient fine-tuning pipeline utilizing Low-Rank Adaptation (LoRA) and 4-bit quantization, we successfully adapted a 4-billion-parameter foundational model (Gemma-3:4B) to perform precise, schema-compliant JSON extraction on administrative impropriety case files. , this research confirms that the general objective was fully achieved: the developed LegalExtract470 this research demonstrates that smaller, locally trained models can achieve state-of-the-art performance in highly specialized domains, significantly reducing the hardware barriers typically associated with Large Language Models.

Our experimental results showed that a pure Supervised Fine-Tuning (SFT) approach, trained on 470 examples, yielded the optimal model. This 4-billion-parameter model, adjusted locally, achieved an impressive F1-score of 0.8250 and a semantic similarity of 0.9021, notably outperforming the significantly larger, computationally expensive Gemma 3:27B 3-shot baseline. This validates the viability of deploying cutting-edge AI in resource-constrained environments. Crucially, this technology holds significant value for the public sector. By automating the reading and structuring of thousands of pages of procurement and lawsuit documents, the proposed approach drastically reduces the manual labor of auditors and prosecutors, thereby boosting institutional efficiency in combating fraud and administrative irregularities.

Despite these strong capabilities, the study highlights important trade-offs and architectural bottlenecks. The model exhibited a slight regression in strict JSON validity (96%) and structural consistency compared to the original few-shot baselines. Furthermore, a granular analysis revealed difficulties in extracting information spread across long documents, particularly for fields requiring broader context such as MUNICÍPIO DE IRREGULARIDADE and FUNÇÃO. This underscores that the 2048-token context window limitation served as a technological barrier during training. Finally, regarding alignment strategies, the hybrid approach (SFT + DAPO) with programmatic reward shaping failed to outperform pure SFT for this specific volume of data, indicating that reinforcement learning techniques require different data scales to yield benefits in structural enforcement. To mitigate this structural fragility in high-temperature regimes, future iterations could scale the fine-tuning dataset to expose the model to a wider variety of formatting contexts, or implement inference-time guardrails such as constrained decoding and token repetition penalties to strictly enforce the target JSON schema.

Moving forward, several promising avenues for future work can be explored to build upon these findings:

1. **Multi-Domain Expansion via Human-LLM-in-the-Loop:** Future research should leverage the hybrid annotation pipeline (Human-LLM in the loop) to expand the current taxonomy beyond administrative impropriety, scaling the approach to encompass other legal spheres, document types, and judicial processes.
2. **Dataset Expansion and RL Optimization:** Expanding the dataset size and diversity could help the model better capture long-context dependencies for difficult fields. A larger dataset (of a few thousand examples) may also provide the necessary distribution for the Reinforcement Learning (DAPO) phase to stabilize and effectively surpass pure SFT, fully utilizing the programmatic reward shaping..
3. **Constrained Decoding:** To mitigate the trade-off between semantic accuracy and structural consistency, future implementations could integrate constrained decoding techniques (e.g., grammar-based generation or strict JSON-schema enforcers) during inference. This would guarantee a 100% JSON validity rate without requiring the model to dedicate parameter capacity to learning syntax.
4. **Retrieval-Augmented Generation (RAG) and Long-Context Handling:** Because legal documents are often lengthy and highly heterogeneous, integrating RAG or chunking strategies could help the model locate dispersed information—such as a defendant’s role introduced pages away from their name—improving accuracy on context-heavy fields.
5. **Cross-Domain Generalization:** Extending the current methodology to other branches of brazilian law (e.g., criminal, civil) or even to other languages could test the robustness and adaptability of the fine-tuning pipeline, potentially leading to a more universally applicable legal information extraction framework.

BIBLIOGRAPHY

- CHENG, T. T. et al. Information extraction from legal documents. In: IEEE. *2009 eighth international symposium on natural language processing*. [S.l.], 2009. p. 157–162. Cited on page 9.
- COLOMBO, P. et al. *SaulLM-7B: A pioneering Large Language Model for Law*. 2024. Disponível em: <<https://arxiv.org/abs/2403.03883>>. Cited 2 times on pages 23 and 25.
- DETTMERS, T. et al. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2022. Cited on page 19.
- DUBEY, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. Disponível em: <<https://arxiv.org/abs/2407.21783>>. Cited on page 17.
- Gemma Team et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. Disponível em: <<https://arxiv.org/abs/2403.08295>>. Cited on page 17.
- GHOLAMI, A. et al. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021. Cited on page 18.
- GODOY, H. *Extract-0: A Specialized Language Model for Document Information Extraction*. 2025. Disponível em: <<https://arxiv.org/abs/2509.22906>>. Cited 3 times on pages 10, 24, and 25.
- HU, E. J. et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. Disponível em: <<https://arxiv.org/abs/2106.09685>>. Cited 2 times on pages 18 and 19.
- HUANG, Q. et al. *Lawyer LLaMA Technical Report*. 2023. Disponível em: <<https://arxiv.org/abs/2305.15062>>. Cited on page 9.
- JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. ed. [S.l.]: Pearson, 2024. Cited on page 14.
- KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Cited on page 19.
- LIU, G. et al. DAPO: Decoupled clip and dynamic sampling policy optimization for large-scale LLM reinforcement learning. *arXiv preprint*, 2025. Cited on page 21.
- LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. Cited on page 19.
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Cited 2 times on pages 14 and 15.
- OPENAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. Cited on page 17.

- OSWALD, J. von et al. *Transformers learn in-context by gradient descent*. 2023. Disponível em: <<https://arxiv.org/abs/2212.07677>>. Cited on page 18.
- OUYANG, L. et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, v. 35, p. 27730–27744, 2022. Cited on page 20.
- PIRES, R. et al. Sabiá: Portuguese large language models. In: _____. *Intelligent Systems*. Springer Nature Switzerland, 2023. p. 226–240. ISBN 9783031453922. Disponível em: <http://dx.doi.org/10.1007/978-3-031-45392-2_15>. Cited on page 23.
- REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. [s.n.], 2019. p. 3982–3992. Disponível em: <<https://arxiv.org/abs/1908.10084>>. Cited on page 16.
- SALTON, G.; WONG, A.; YANG, C.-S. A vector space model for automatic indexing. *Commun. ACM*, v. 18, p. 613–620, 1975. Disponível em: <<https://api.semanticscholar.org/CorpusID:6473756>>. Cited on page 14.
- SCHULMAN, J. et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. Cited on page 20.
- SHANNON, C. E. Prediction and entropy of printed english. *Bell System Technical Journal*, v. 30, n. 1, p. 50–64, 1951. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1951.tb01366.x>>. Cited on page 14.
- SHAO, Z. et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Cited on page 21.
- SUTTON, R. S.; BARTO, A. G. *Reinforcement learning: An introduction*. [S.l.]: MIT press, 2018. Cited on page 20.
- TEAM, G. et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. Cited on page 17.
- TOUVRON, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. Cited on page 17.
- TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, AI Access Foundation, v. 37, p. 141–188, fev. 2010. ISSN 1076-9757. Disponível em: <<http://dx.doi.org/10.1613/jair.2934>>. Cited on page 14.
- VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017. Cited 3 times on pages 9, 14, and 15.
- WAN, Z. et al. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on Chinese legal domain. In: KU, L.-W.; MARTINS, A.; SRIKUMAR, V. (Ed.). *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 5030–5041. Disponível em: <<https://aclanthology.org/2024.findings-acl.299/>>. Cited on page 9.

WANG, X. et al. *InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction*. 2023. Disponível em: <<https://arxiv.org/abs/2304.08085>>. Cited 2 times on pages 24 and 25.

WILLARD, B. T.; LOUF, R. *Efficient Guided Generation for Large Language Models*. 2023. Disponível em: <<https://arxiv.org/abs/2307.09702>>. Cited on page 24.

YANG, A. et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. Disponível em: <<https://arxiv.org/abs/2407.10671>>. Cited on page 17.

ZHAO, W. X. et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. Cited on page 17.

A SBC PAPER

This appendix contains the paper produced in SBC format that refers to the present work.

Parameter-Efficient Fine-Tuning and Alignment of Quantized LLMs for Strict Schema Legal Extraction

João Pedro P. Resmer¹, Carina Friedrich Dorneles¹, Vitoria Soares dos Santos¹

¹ Depto. de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88040-900 – Florianópolis – SC – Brazil

joao.pedro.perez.resmer@grad.ufsc.br
carina.dorneles@ufsc.br
vitoria.s@posgrad.ufsc.br

Abstract. *Large Language Models (LLMs) have emerged as very powerful tools for multiple tasks over the past years. Their ability to approximate human performance in a variety of fields through next-token prediction has attracted much attention inside and outside the academic community. Given their ability to represent semantics in high-dimensional vector spaces using embedding models and process them using self-attention, this work proposes two fine-tuning approaches using Low-Rank Adaptation (LoRA) and quantization to leverage the great capabilities of LLMs for a labor-intensive task, data extraction from legal documents, while minimizing the hardware requirements for training a 4-billion-parameter model.*

1. Introduction

As digitization spreads across various fields worldwide, Brazil is no exception. In fact, the country currently has an extensive array of digitized documents and procedures available. A notable example is the national ID, which, since 2022, has featured a valid online version accessible to anyone who has had their document issued since then. The judicial system has followed this push toward digitized procedures. This shift has significantly enlarged the production and storage of legal documents, making them a vast source of information for legal professionals such as lawyers, judges, and prosecutors. However, this creates a new challenge: extracting useful information from a massive database of documents that are semantically complex and extremely diverse in length, content, context, and structure to then formatting it into a schema adherent json object is a highly difficult and labor-intensive task.

Specifically, between 2020 and 2024, Brazil registered over 205,000 procedures related solely to bidding processes. Each of these procedures requires multiple documents to be issued, creating a massive influx of information into the Brazilian government’s systems. Consequently, manually reviewing these documents to search for useful information applicable to a specific subset of judicial procedures becomes an extremely time-consuming task.

Given these challenges, it is natural to turn to computational solutions. Natural Language Processing (NLP) has long been an area of interest, and various methods have been developed to extract specific sequences of characters from text. The scientific community has utilized many different approaches, ranging from pattern matching

[Cheng et al. 2009] to regular expression-based methods, achieving varying levels of performance. While consistent and structured data with clear formatting rules can be efficiently retrieved, extracting relevant information from diverse, unstructured excerpts has remained challenging. This remained the case until the introduction of the Transformer architecture [Vaswani et al. 2017], which significantly advanced the field of natural language processing.

This new architecture allowed deep learning models to effectively process semantics in text embeddings and achieve unprecedented text generation capabilities. However, certain difficulties persisted. As noted by [Huang et al. 2023], these models often lack domain-specific knowledge and struggle to apply general knowledge to solve domain-related tasks. Despite these challenges, recent studies have demonstrated remarkable performance in information extraction tasks. For instance, [Wan et al. 2024] proposed a zero-shot, three-step pipeline capable of enhancing the performance of GPT-4 on Chinese legal documents, while [Godoy 2025] employed fine-tuning techniques to improve performance on a more diverse dataset.

The present work focuses on addressing these gaps by training a model to achieve state-of-the-art performance on a legal data extraction task. While commercial, cloud-based Large Language Models are powerful, their application in the legal domain is often restricted by strict document confidentiality requirements and prohibitive inference costs at scale. Therefore, this research emphasizes a local processing approach. Using quantization and Low-Rank Adaptation (LoRA), we locally train a 4-billion-parameter model (Gemma 3). The model is fine-tuned on a custom dataset consisting of 100 legal documents (50 hand-annotated, and 50 automatically annotated with Llama 3.3 using a 10-shot prompting strategy and subsequently human-revised) related to administrative impropriety.

The task of extracting structured data from legal documents is highly complex, requiring the model to comprehend intricate legal semantics and output a strictly schema-adherent JSON object. Specifically, the structured output must accurately capture information across three core fronts: *dados da licitação* (bidding data), *dados do processo* (process data), and *dados dos requeridos* (defendants' data). To comply with a 2048-token limit, each document was segmented into multiple instances, with each instance comprising a document excerpt paired with its corresponding structured JSON object. As a result, the annotation process produced over 570 distinct examples. To ensure rigorous evaluation, 100 of these examples were reserved exclusively as a test set, while the remainder constituted the training data. Additionally, we employ a custom loss function during fine-tuning that leverages cosine similarity to more accurately score outputs that capture the correct underlying semantic information. In order to better understand the model's capabilities to adapt to the task under different training settings, we also employ a second approach combining Supervised Fine-Tuning with Reinforcement Learning. Dynamic Sampling Policy Optimization (DAPO) with a programmatic reward to directly optimize the policy against structural adherence and semantic precision, subsequently comparing the results in order to produce a final model (LegalExtract470).

2. Related Work

The task of extracting structured data from Brazilian legal documents using Large Language Models (LLMs) lies at the intersection of multiple areas of Natural Language Processing (NLP) research. In order to contextualize this work and its contributions, the present section reviews relevant literature across three primary domains. First, it explores the adaptation of LLMs to the legal domain. Second, it analyzes work on specialized models designed for document-level information extraction. Lastly, it focuses on instruction tuning for structured extraction. By examining these foundational works, this section establishes the current state-of-the-art on similar tasks and highlights research gaps.

2.1. Domain-Specific LLMs for Law

Domain specific adaptation for highly technical fields, such as law, have been a source of recent advancements. Amongst them, a notable contribution is SaulLM-7B [Colombo et al. 2024], a pioneering model designed for legal text. The model is an adaptation of Mistral 7B trained on an extensive legal corpus in order to comprehend legal vocabulary and reasoning. However, while SaulLM establishes the viability of such models for legal reasoning, its efforts were focused on on general legal comprehension in the English language. Additionally, the model’s training and evaluation is aimed towards more classical Legal NLP tasks, such as summarization. Consequently, the ability of smaller foundational models, such as Gemma-3:4B, to learn the syntactic and lexical nuances of Brazilian legal documents is still not widely explored.

2.2. Local Language Adaptation for Legal Text

While domain-specific models like SaulLM-7B demonstrate the power of continuous pre-training for legal language, they remain fundamentally limited by their focus on the English language and Anglo-Saxon legal systems. Within the Brazilian legal context, recent efforts have focused on adapting Large Language Models to Portuguese. For instance, the Sabiá family of models [Pires et al. 2023] demonstrated that careful curation of Portuguese corpora yields competitive performance on local benchmarks without requiring massive scale.

However, a persistent research gap remains at the intersection of local language adaptation and structural extraction tasks. While models like Sabiá or Llama-BR excel at general text understanding and legal classification, they are rarely evaluated or optimized for generating strict, deeply nested structures—such as JSON schemas—under local hardware constraints. This work addresses this limitation by taking Gemma 3, a highly efficient, multilingual small language model, and explicitly aligning it for structured extraction within the complex domain of Brazilian Administrative Misconduct actions.

2.3. LLMs for Document Information Extraction

Although domain adaptation may improve a model’s capabilities for legal domain tasks in general, extracting structured information directly from unstructured documents requires specialized training. [Godoy 2025] addressed this challenge with Extract-0, a specialized language model trained for document information extraction. Extract-0 highlights the efficacy of specializing language models for document parsing as an alternative to traditional pipeline-based NLP extraction. However, applying such methodologies to often ambiguous and more heterogeneous text, while producing highly structured, schema-dependent

outputs—like strictly formatted JSONs required in legal tech workflows—presents distinct challenges that often require customized optimization strategies.

2.4. Multi-Task Instruction Tuning for Unified Extraction

To enable models to generalize across varied extraction tasks, researchers have increasingly relied on instruction tuning. InstructUIE, introduced by [Wang et al. 2023], is a comprehensive framework that utilizes multi-task instruction tuning to create a unified information extraction system. By framing various extraction tasks (such as entity, relation, and event extraction) as natural language generation tasks, InstructUIE proved that seq-to-seq models could handle complex, multi-faceted extraction instructions simultaneously. InstructUIE serves as a foundational proof-of-concept that instruction tuning is a highly effective method for teaching models to map unstructured text to structured outputs. However, their reliance on standard Supervised Fine-Tuning (SFT) utilizing standard Cross-Entropy (CE) loss may not be sufficient when rigid adherence to complex data schemas is strictly required.

2.5. Structured Generation and Constrained Decoding

A critical challenge in leveraging LLMs for Information Extraction (IE) is ensuring that the generated output strictly adheres to a predefined schema (e.g., JSON). Traditional IE approaches rely heavily on fine-tuning to teach the model both the semantic extraction task and the syntactic rules of the target format. To mitigate formatting failures, recent literature has proposed mechanisms such as constrained decoding. A prominent framework is Outlines [Willard and Louf 2023], which enforces formal grammar constraints directly at the token sampling level during inference, effectively guaranteeing 100% syntactical validity without altering the model’s internal weights.

In this study, we choose not to employ inference-time grammar constraints during the main evaluation phases. This decision allows us to isolate and measure the raw alignment capability of the SFT and DAPO training paradigms. By analyzing the model’s unconstrained output, such as the 96% JSON validity rate observed, we can accurately assess how effectively the custom loss functions and reinforcement learning steps imbue the model with structural discipline, leaving constrained decoding frameworks like Outlines as an orthogonal, complementary optimization for production deployment.

3. Methodology

This section details the proposed fine-tuning approach to extract structured information from legal documents. The pipeline encompasses data annotation, model initialization via quantization, Low-Rank Adaptation (LoRA), and two distinct training paradigms: a Supervised Fine-Tuning (SFT) with a custom multi-objective loss, and a two-stage approach leveraging Reinforcement Learning (RL).

3.1. Task Formulation and Document Structure

The extraction task is formulated as a structured sequence-to-sequence problem. Given a raw text excerpt from a legal document—specifically case files of administrative impropriety—the model must generate a strictly formatted JSON object.

The source documents originate from different counties and follow varying procedural norms, resulting in significant structural variations. Nevertheless, common elements are present across the corpus. Typically, documents begin with a header indicating the county of origin and an identification block containing the civil case ID or the SIG. This is usually followed by a list of defendants, including their names, descriptions, and roles. The concluding sections of the documents contain the claims, which often reference the type of administrative impropriety, the claim values, the date, and the prosecutor responsible.

The expected JSON output consists of three primary root keys: "DADOS PROCESSO", "DADOS DOS REQUERIDOS", and "DADOS LICITAÇÃO", which map to structured data defined in Tables 1 and 2. The model must map varied textual formulations and legal jargon into this rigid schema. If the text excerpt lacks the necessary information for a specific field, the model is instructed to output the string "Null". This semantic gap fundamentally motivates the implementation of a specialized loss function and reward shaping mechanisms.

Table 1. JSON schema definition for case/case file extraction.

Key	Value Definition
INQUÉRITO ASSOCIADO	ID of the civil case
SIG (RELACIONADO AO INQUÉRITO)	SIG ID tied to the civil case
COMARCA DE AJUIZAMENTO DA AÇÃO	County where the case was listed
DATA DE PUBLICAÇÃO DO PROCESSO	Date when the case was filed
VALOR DA CAUSA	Financial value of the claim
TESTEMUNHAS	List of witnesses
TIPO DE IMPROBIDADE	Type(s) of administrative impropriety
PROMOTOR	Prosecutor(s)

Table 2. JSON schema definition for public bidding extraction.

Key	Value Definition
NÚMERO PROCESSO LICITATÓRIO	ID of the overall bidding process
NÚMERO EDITAL (PROCESSO LICITATÓRIO)	ID of the specific bidding notice
DATA DE PUBLICAÇÃO DO EDITAL	Registration date of the bidding
VALOR DO SERVIÇO	Cost associated with the contract
SUPERFATURAMENTO	Estimated cost of overbilling
TIPO DE PRODUTO-SERVIÇO	Goods and services contracted

3.2. Dataset Annotation

The raw data was provided by Projeto Ceos, an initiative from the Federal University of Santa Catarina (UFSC) in partnership with the Public Prosecutor's Office of Santa Catarina (MPSC). From a pool of over 5,000 documents, 100 were randomly selected to compose the dataset. The first 50 documents were manually annotated into the predefined JSON format. The remaining 50 were pre-annotated using Llama 3.3 (70B parameters) via 10-shot prompting. Every pre-annotated instance was then manually verified and corrected through careful cross-referencing to ensure a high-quality dataset.

3.3. Model Initialization and LoRA Configuration

Training was performed on an NVIDIA H100 GPU using Hugging Face libraries. Gemma 3, a 4-billion-parameter foundation model, was selected due to its strong baseline performance on complex extraction tasks.

To enable training within computational constraints, the base model weights were quantized to 4-bit NormalFloat (NF4) precision, with double quantization enabled and a 16-bit floating-point (FP16) compute data type during the forward pass. Parameter-Efficient Fine-Tuning was implemented via Low-Rank Adaptation (LoRA) configured for a Causal Language Modeling task. The adapter matrices were initialized with a rank of $R = 16$, an alpha scaling factor of $\alpha = 32$, and a dropout rate of 10%. Crucially, LoRA adapters were injected comprehensively across the Transformer architecture, targeting all self-attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`) and Multi-Layer Perceptron components (`up_proj`, `down_proj`, and `gate_proj`) to maximize model plasticity.

3.4. Supervised Fine-Tuning (SFT) and Custom Loss

The models were evaluated iteratively on subsets up to 470 examples, holding out 20% for validation. Due to hardware limitations, the batch size was set to 4 with 4 gradient accumulation steps (effective batch size of 16), and the maximum sequence length was capped at 2,048 tokens. The Paged 8-bit AdamW optimizer was utilized with a learning rate of 3×10^{-4} and a cosine scheduler over 8 epochs, employing an early stopping patience of 3 epochs.

Standard Cross-Entropy penalizes outputs that are semantically equivalent but lexically different from the ground truth and fails to enforce hierarchical JSON structures. To address this, a multi-objective loss function was introduced, defined in Equation 1:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{tree}}(t)\mathcal{L}_{\text{tree}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{struct}}\mathcal{L}_{\text{struct}} \quad (1)$$

The components comprise:

1. **Cross-Entropy ($\mathcal{L}_{\text{ce}}$):** The standard autoregressive token-level negative log-likelihood.
2. **Semantic Loss ($\mathcal{L}_{\text{sem}}$):** Computes the cosine distance between the mean-pooled sequence embeddings of the expected predicted tokens and the ground-truth representations, penalizing semantic divergence.
3. **Tree Alignment Loss ($\mathcal{L}_{\text{tree}}$):** Utilizes Optimal Transport (via Sinkhorn-Knopp algorithm) to match the hierarchical length and structure of the target JSON chunks. Its weight $\lambda_{\text{tree}}(t)$ is dynamically warmed up to prevent early training instability.
4. **Structural Loss ($\mathcal{L}_{\text{struct}}$):** Enforces JSON schema generation by penalizing low predicted probabilities for fundamental JSON syntax characters (e.g., braces, brackets, colons).

Empirically, static scalar weights were set to $\lambda_{\text{ce}} = 0.65$, $\lambda_{\text{sem}} = 0.55$, and $\lambda_{\text{struct}} = 0.30$, with λ_{tree} warming up to 0.50.

3.5. Reinforcement Learning Phase (Two-Stage Approach)

As an alternative to purely custom-loss SFT, a two-stage training paradigm was evaluated. The model first undergoes standard Cross-Entropy SFT to acquire a baseline understanding of extraction and formatting. Subsequently, Reinforcement Learning (RL) is applied using the Dynamic Sampling Policy Optimization (DAPO) algorithm.

The environment generates a group of $G = 4$ distinct completions for each input, and the algorithm calculates the training advantage by normalizing the programmatic reward scores. To maintain linguistic coherence and avoid reward hacking, a Kullback-Leibler (KL) divergence penalty ($\beta = 8.5$) was enforced. This stage utilized a learning rate of 1×10^{-5} and 16 gradient accumulation steps.

Since JSON extraction possesses objective success criteria, a programmatic reward function was engineered. The final scalar reward (R_{total}) linearly combines two factors (Equation 3): a binary **Format Reward** (R_{format}) that checks for valid JSON parsing, and a **Semantic Reward** (R_{semantic}). The semantic reward maps the flattened predictions and ground truth strings into a dense vector space using a lightweight Sentence Transformer (all-MiniLM-L6-v2), returning a normalized cosine similarity (Equation 2).

$$R_{\text{semantic}}(y_{\text{pred}}, y_{\text{gt}}) = \frac{\text{sim}_{\cos}(E(y_{\text{pred}}), E(y_{\text{gt}})) + 1}{2} \quad (2)$$

$$R_{\text{total}} = \text{clip}(w_1 R_{\text{format}} + w_2 R_{\text{semantic}}, 0, 1) \quad (3)$$

3.6. Generative AI Usage

Generative AI models served as instrumental tools across the project’s lifecycle. Llama 3.3 assisted in the initial dataset pre-annotation (subsequently human-verified) and supported code generation and debugging. The Gemma embedding model and the lightweight Sentence Transformer were central to computing semantic representations in both the Custom Loss function and the RL Programmatic Reward Shaping. Additionally, LLMs provided writing assistance for text verification, significantly accelerating development and enhancing the structural clarity of the pipeline.

4. Experimental Design

The proposed experiments aim to evaluate the performance of the fine-tuned model on the structured legal extraction task. The evaluation is designed around four primary objectives:

- **Baseline Comparison (Goal 1):** Compare the extraction accuracy and structural consistency of the proposed model against the original zero-shot and few-shot Gemma 3 baselines.
- **Generalization Assessment (Goal 2):** Evaluate model performance on unseen data using a held-out test set of 100 annotated documents.
- **Multi-Dimensional Evaluation (Goal 3):** Assess the generations across multiple dimensions, ranging from structural integrity to lexical precision.
- **Robustness Assessment (Goal 4):** Analyze robustness across varied data formats (e.g., highly structured dates versus unstructured witness lists) through a granular token-level analysis.

4.1. Evaluation Metrics

To capture distinct dimensions of model performance, we employ six evaluation metrics:

- **Tree Edit Distance (TED):** Evaluates the structural and hierarchical integrity of the generated JSON. Both the predicted and ground-truth JSON objects are converted into tree structures, with keys sorted to ensure order invariance. The All-Path Tree Edit Distance (APTED) algorithm computes the minimum node operations (insertions, deletions, and substitutions) required to align the prediction tree with the reference. This raw distance is normalized by the size of the ground-truth object to produce a structural similarity score bounded within $[0, 1]$.
- **F1-Score (F_1):** Measures the harmonic mean of precision and recall at the key-value field level. The predicted and ground-truth JSON objects are flattened into sets of unique $(\text{path}, \text{value})$ string pairs, where the path ignores ordering. The micro F1-score is calculated based on exact matches within these sets to assess overall extraction accuracy.
- **Validity Rate (v_r):** Represents the proportion of generations that yield a well-formed JSON string parseable by a standard decoder, after isolating the content between the first $\{$ and last $\}$ characters:

$$v_r = \frac{\text{valid}}{\text{total}} \quad (4)$$

- **Semantic Similarity (SSim):** Quantifies semantic equivalence independent of structural formatting, accounting for valid paraphrasing. Extracted values from the prediction and target JSONs are encoded into dense vectors using a Sentence Transformer model. A cosine similarity matrix is calculated between these sets, and the final score is computed by averaging the maximum similarity scores found for each predicted value against all ground-truth values.
- **Consistency (Consis):** Measures the deterministic stability of the model’s outputs across multiple independent generation cycles. Raw outputs from repeated inference runs are cleaned, parsed, and compared pair-wise. For each unique pair of valid structures, a dedicated F1-score measures the exact overlap of extracted keys and values. The final consistency score is the arithmetic mean of these pairwise scores, indicating robustness against prompt ambiguity.
- **Token Accuracy:** Calculates the ratio of exactly matched tokens between the generated and ground-truth JSON outputs, providing a fine-grained measure of lexical correctness.

4.2. Prompting Strategies

To establish robust baselines and guide model extraction, we designed two distinct prompting configurations in Portuguese to match the language of the target legal documents:

1. **Zero-Shot Baseline Prompt:** Defines the system instructions and the expected output schema. The prompt instructs the model to extract information from a target text delimited by `[alvo]` and format the output as a valid JSON object after a `[saida]` tag. The expected schema is structured into three main blocks:

DADOS PROCESSO (containing metadata like case IDs, dates, and court locations), DADOS DOS REQUERIDOS (a nested list containing the names, functions, and identification numbers of the defendants), and DADOS LICITAÇÃO (detailing bidding IDs, values, and modalities). Specific guidelines instruct the model to map the TIPO DE IMPROBIDADE field to relevant articles of the Brazilian Administrative Impropriety Law (Law No. 8.429/92, Arts. 9, 10, or 11) and fill missing fields with the string "Null".

2. **Few-Shot Baseline Prompt:** Extends the system instructions and schema specifications defined in the first prompt by incorporating in-context learning. This version includes three diverse, hand-annotated input-output examples drawn from the training set, allowing the baseline models to better align with the expected lexical density and output structure.

4.3. Validation and Model Selection

Validation was performed using the metrics and prompts described in the evaluation setup. A validation set, comprising 20% of the annotated examples, was used to monitor the training progression and mitigate overfitting. Model checkpoints and hyperparameters were selected based on the highest performance across three main validation metrics: the F1-score, Semantic Similarity, and the Validity Rate. The selected final checkpoint was subsequently evaluated on the held-out test set.

5. Results

The results of the experiments conducted in this study are presented in this chapter. The performance of the fine-tuned models is evaluated using a set of metrics that assess both the semantic accuracy and structural consistency of the extracted JSON data. The evaluation is performed on a held-out test set, which was not used during the training or validation phases.

5.1. Comparison of The Two Tuning Approaches

Table 3. Tuned models' performance comparison on the structured legal data extraction task.

Model	TED	F1-Score	SSim	Valid Rate	Consis
Pure SFT (100 examples)	0.7416	0.6246	0.6842	0.7500	0.4573
Pure SFT (200 examples)	0.7892	0.6486	0.7296	0.8000	0.6187
Pure SFT (300 examples)	0.7897	0.6346	0.7318	0.8000	0.5773
Pure SFT (400 examples)	0.8763	0.7362	0.8174	0.8875	0.6509
Pure SFT (470 examples)	0.9170	0.8280	0.9114	0.9255	0.7222
SFT + RL	0.9287	0.8095	0.8856	0.9362	0.6920

As seen in Table 3, the pure SFT approach with 470 examples achieved the best performance across multiple metrics on the validation set. The SFT + RL approach did not outperform the pure SFT method in this validation phase, indicating that the additional complexity of reinforcement learning in this restricted training set may not have provided a significant benefit for this specific task and dataset size. Based on these results, the pure SFT model trained with 470 examples was selected as the final model checkpoint for evaluation on the held-out test set.

5.2. Test Set Evaluation

This section presents the performance evaluation of the baseline models and the fine-tuned model, `LegalExtract470`, on the held-out test set. The quantitative results across the six evaluation metrics are summarized in Table 4.

Table 4. Global performance comparison of baseline models and the proposed `LegalExtract470` model on the structured legal data extraction task.

Model	TED	F1-Score	SSim	Valid Rate	Consis	Token Acc.
Qwen2.5:3B (3-shot)	0.9784	0.6807	0.8203	1.0000	0.7243	0.0097
Gemma3:27B (3-shot)	0.9859	0.7947	0.9229	1.0000	0.9821	0.4969
Gemma3:4B (3-shot)	0.9733	0.7488	0.8909	0.9900	0.9353	0.3862
Gemma3:4B (raw)	0.9641	0.4548	0.8273	1.0000	0.8654	0.1559
<code>LegalExtract470</code>	0.9516	0.8250	0.9021	0.9600	0.7254	0.4173

First, we observe the impact of In-Context Learning on the baseline foundation models. Comparing the zero-shot baseline `Gemma3:4B (raw)` to its three-shot counterpart, `Gemma3:4B (3-shot)`, reveals a substantial performance improvement across nearly every metric, with the most notable increases occurring in F1-score and token accuracy. Nonetheless, the smaller 3B parameter model (`Qwen2.5:3B`) struggled with extraction accuracy, producing a token match score of only 0.0097, despite its perfect valid JSON rate.

Second, the proposed fine-tuned model, `LegalExtract470`, demonstrates strong performance in factual retrieval. It achieves the highest F1-score (0.8250) among all evaluated models, including the significantly larger `Gemma3:27B (3-shot)` baseline (0.7947). Our model also produces a competitive semantic similarity score of 0.9021 and an token match accuracy of 0.4173. This indicates that across all valid JSON structures generated during testing, over 41% of the individual tokens perfectly matched their corresponding target tokens. Consequently, `LegalExtract470` outperforming other baseline models of comparable scale at a lower computational cost.

However, parameter-efficient fine-tuning introduced a notable trade-off between semantic extraction capability and structural consistency. Because the consistency evaluation employs a higher sampling temperature, the fine-tuned model becomes susceptible to autoregressive text degeneration. When a low-probability or suboptimal token is sampled early in the generation sequence, the model enters a self-reinforcing repetition loop, effectively losing its contextual footing and continuously generating degenerate formatting. Consequently, while `LegalExtract470` excels at identifying and correctly retrieving specific legal key-value pairs, it exhibits a slight regression in both structural consistency and JSON validity (dropping to 96%) when compared to the three-shot baselines.

5.2.1. Per-Field Granular Analysis

To better understand the practical performance of the models within the Brazilian administrative impropriety domain, a granular analysis of extraction accuracy at the token level was conducted. For that a semantic accuracy metric was used (an answer is deemed correct in this case if its cosine similarity to the expected value for the field is

over 0.80) for each field in the DADOS LICITAÇÃO, DADOS PROCESSO, and DADOS DOS REQUERIDOS sections. The study reveals several key architectural behaviors:

1. **Systematically mastered fields:** Highly structured metadata, such as dates (DATA DE PUBLICAÇÃO DO PROCESSO, DATA DE PUBLICAÇÃO DO EDITAL), case IDs (INQUÉRITO ASSOCIADO, SIG), and bidding modalities (MODALIDADE), are reliably extracted across almost all models, as represented by the bright yellow bands. These fields generally follow strict formatting rules for legal headers, making them easily identifiable.
2. **Complex, unstructured fields:** A key finding concerns the performance on the DADOS PROCESSO.TESTEMUNHAS (witnesses) and DADOS PROCESSO.TIPO DE IMPROBIDADE (type of impropriety) fields. Although zero-shot and few-shot baselines largely fail to extract witness lists from legal prose and differentiate them from DADOS DOS REQUERIDOS entries, LegalExtract470 achieves an exceptionally high token accuracy (represented by the bright yellow highlight in its column). This indicates that fine-tuning successfully enabled the model to learn a more nuanced approach to this extraction task.
3. **Inherent dataset bottlenecks:** Finally, fields like DADOS PROCESSO.MUNICÍPIO DE IRREGULARIDADE or the more complex nested structure in DADOS DOS REQUERIDOS represent a clear bottleneck in the performance of all models. This limitation may stem from the inherent nature of these fields. For instance, DADOS PROCESSO.MUNICÍPIO DE IRREGULARIDADE is easily confused with COMARCA DE AJUIZAMENTO DA AÇÃO (county of filing) without a careful analysis of the entire document's context. Similarly, DADOS DOS REQUERIDOS.FUNÇÃO (respondent's role) is a complex field because it may not be present in the paragraph where the respondent is first introduced. Furthermore, its correct interpretation requires a broader understanding of the illegalities described throughout the entire document, which is never available to the model as we are constrained by the sequence length during training. Additionally, this outcome may stem from a lack of sufficient training examples for specific fields, particularly given that legal documents can be highly heterogeneous.

6. Conclusion and Future Work

This work aimed to address the challenging and labor-intensive task of structured data extraction from heterogeneous Brazilian legal documents. By proposing a parameter-efficient fine-tuning pipeline utilizing Low-Rank Adaptation (LoRA) and 4-bit quantization, we successfully adapted a 4-billion-parameter foundational model (Gemma-3:4B) to perform precise, schema-compliant JSON extraction on administrative impropriety case files. This research confirms that the general objective was fully achieved: the developed LegalExtract470 this research demonstrates that smaller, locally trained models can achieve state-of-the-art performance in highly specialized domains, significantly reducing the hardware barriers typically associated with Large Language Models.

Our experimental results showed that a pure Supervised Fine-Tuning (SFT) approach, trained on 470 examples, yielded the optimal model. This 4-billion-parameter

model, adjusted locally, achieved an impressive F1-score of 0.8250 and a semantic similarity of 0.9021, notably matching the performance of the significantly larger, computationally expensive Gemma 3:27B 3-shot baseline. This validates the viability of deploying cutting-edge AI in resource-constrained environments. Crucially, this technology holds significant value for the public sector. By automating the reading and structuring of thousands of pages of procurement and lawsuit documents, the proposed approach drastically reduces the manual labor of auditors and prosecutors, thereby boosting institutional efficiency in combating fraud and administrative irregularities.

Despite these strong capabilities, the study highlights important trade-offs and architectural bottlenecks. The model exhibited a slight regression in strict JSON validity (96%) and structural consistency compared to the original few-shot baselines. Furthermore, a granular analysis revealed difficulties in extracting information spread across long documents, particularly for fields requiring broader context such as `MUNICÍPIO DE IRREGULARIDADE` and `FUNÇÃO`. This underscores that the 2048-token context window limitation served as a technological barrier during training. Finally, regarding alignment strategies, the hybrid approach (SFT + DAPO) with programmatic reward shaping failed to outperform pure SFT for this specific volume of data, indicating that reinforcement learning techniques require different data scales to yield benefits in structural enforcement. To mitigate this structural fragility in high-temperature regimes, future iterations could scale the fine-tuning dataset to expose the model to a wider variety of formatting contexts, or implement inference-time guardrails such as constrained decoding and token repetition penalties to strictly enforce the target JSON schema.

Moving forward, several promising avenues for future work can be explored to build upon these findings:

1. **Multi-Domain Expansion via Human-LLM-in-the-Loop:** Future research should leverage the hybrid annotation pipeline (Human-LLM in the loop) to expand the current taxonomy beyond administrative impropriety, scaling the approach to encompass other legal spheres, document types, and judicial processes.
2. **Dataset Expansion and RL Optimization:** Expanding the dataset size and diversity could help the model better capture long-context dependencies for difficult fields. A larger dataset (of a few thousand examples) may also provide the necessary distribution for the Reinforcement Learning (DAPO) phase to stabilize and effectively surpass pure SFT, fully utilizing the programmatic reward shaping..
3. **Constrained Decoding:** To mitigate the trade-off between semantic accuracy and structural consistency, future implementations could integrate constrained decoding techniques (e.g., grammar-based generation or strict JSON-schema enforcers) during inference. This would guarantee a 100% JSON validity rate without requiring the model to dedicate parameter capacity to learning syntax.
4. **Retrieval-Augmented Generation (RAG) and Long-Context Handling:** Because legal documents are often lengthy and highly heterogeneous, integrating RAG or chunking strategies could help the model locate dispersed information—such as a defendant’s role introduced pages away from their name—improving accuracy on context-heavy fields.
5. **Cross-Domain Generalization:** Extending the current methodology to other branches of Brazilian law (e.g., criminal, civil) or even to other languages could

test the robustness and adaptability of the fine-tuning pipeline, potentially leading to a more universally applicable legal information extraction framework.

References

- Cheng, T. T., Cua, J. L., Tan, M. D., Yao, K. G., and Roxas, R. E. (2009). Information extraction from legal documents. In *2009 eighth international symposium on natural language processing*, pages 157–162. IEEE.
- Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., and Desa, M. (2024). Saullm-7b: A pioneering large language model for law.
- Godoy, H. (2025). Extract-0: A specialized language model for document information extraction.
- Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., and Feng, Y. (2023). Lawyer llama technical report.
- Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- 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. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5030–5041, Bangkok, Thailand. Association for Computational Linguistics.
- Wang, X., Zhou, W., Zu, C., Xia, H., Chen, T., Zhang, Y., Zheng, R., Ye, J., Zhang, Q., Gui, T., Kang, J., Yang, J., Li, S., and Du, C. (2023). Instructuie: Multi-task instruction tuning for unified information extraction.
- Willard, B. T. and Louf, R. (2023). Efficient guided generation for large language models.

B SOURCE CODE AND DATA

While the data cannot be shared due to legal restrictions, the source code for the experiments and the fine-tuned model weights are available at <https://codigos.ufsc.br/joao.pedro.perez.resmer/tcc>. The repository contains the code for training and evaluating the models, as well as the list dependencies necessary to run the code on `requirements.txt`, version compatibility for the dependencies may be slightly different depending on setup. Additionally, `pyproject.toml` and `uv.lock` were produced outside of the original environment and may or may not be fully compatible with the original setup.