Democratizing Legal Analytics: Resource-Efficient Information Extraction for Brazilian Case Law*

Rodrigo Filippi Dornelles
Hertie School / Berlin, Germany
r.dornelles@alumni.hertie-school.org

Abstract

Legal systems produce large volumes of high-stakes decisions in unstructured natural language, making large-scale empirical analysis costly, difficult to reproduce, and unevenly accessible. This bottleneck is especially acute for legal analytics and policy evaluation in low-resource languages such as Portuguese. To address it, we present a resource-efficient pipeline for information extraction from Brazilian criminal case law that reuses a legacy dataset to fine-tune open-weight LLMs with Q-LoRA. Operating in a small-data setting and using schema-constrained JSON generation, the pipeline extracts 47 legal variables spanning charges, evidence, and sentencing outcome. In held-out evaluation, a fine-tuned Phi-4 (14B) model achieves 92.8% accuracy and 0.826 macro-F1, approaching proprietary baselines while retaining the cost and privacy benefits of local deployment. We then use the extracted data in a case study of the short-term effects of a recent Brazilian Supreme Court ruling on drug decriminalization, finding no statistically significant change in trafficking-conviction rates ($p \geq 0.05$), a pattern consistent with short-run institutional inertia. More broadly, the paper contributes a reproducible framework for legal NLP and shows how legacy empirical datasets can support scalable legal analytics under severe resource constraints.

1 Introduction

The Brazilian Judiciary is one of the largest in the world, with almost 300 million digital lawsuits (Conselho Nacional de Justiça, 2025). This volume renders manual analysis of case law and legislative impact virtually impossible. Although Jurimetrics — the statistical study of law (Nunes, 2016) — has emerged as a solution, it faces significant barriers due to the unstructured nature of legal data and the scarcity of models capable of comprehending the specific nuances of the Portuguese Brazilian legal jargon. In contrast, traditional Natural Language Processing (NLP) techniques often fail to capture the complexities of legal syntax and context. Building on the architectural framework initially proposed in Dornelles (2025a,b), we validate a pipeline that uses fine-tuned open-source Large Language Models (LLMs)[1] to automate the extraction of legal variables.

In June 2024, the Brazilian Supreme Federal Court (STF) ruled on the partial decriminalization of cannabis possession for personal use (RE 635.659) (Supremo Tribunal Federal, 2024). To assess the effects of this ruling without costly expert annotation, we analyzed a sample of lower-court decisions using LLMs to extract structured information from judicial sentences. We repurposed prior empirical legal data to fine-tune models for this extraction task and then applied the fine-tuned model to generate the variables required to answer our research question (Dornelles, 2025a).

We apply this methodology to a 454-case sample (Dornelles, 2025b) to investigate our substantive research question: *Has this binding STF precedent effectively changed judicial outcomes in São Paulo's first-instance courts*?

Contributions[2]: this work establishes a methodological blueprint by: (1) using legacy data (Lacerda e Silva, 2021) to bootstrap modern LLMs; (2) demonstrating resource-efficient adaptation, where smaller open-source models fine-tuned with Q-LoRA match proprietary baselines; and (3) coupling NLP with classical statistics as reliable jurimetric instrument for evaluating judicial policy.

*This work is an adapted and extended version of research originally conducted for the author's Master's thesis at the Hertie School and MBA at the University of Sao Paulo.

[1] For brevity, we use Large Language Models (LLMs) as an umbrella term that also encompasses parameter-efficient Small Language Models (SLMs).

[2] Code available at https://github.com/rfdornelles/propor2026-legal-nlp

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1011–1020

2 Related Work

Legal NLP in Portuguese: While computational methods have increasingly transformed legal analysis globally (Frankenreiter and Livermore, 2020), their application remains underutilized in the Brazilian context, where empirical legal research has traditionally relied on manual annotation (Costa et al., 2011; Lacerda e Silva, 2021; IPEA and Ministério da Justiça, 2023) or classical approaches such as regular expressions. These methods, however, are often limited by the complexity of legal reasoning. While recent LLMs specialized in Portuguese — such as Sabiá Family (Pires et al., 2023; Almeida et al., 2024; Abonizio et al., 2025), Tucano (Corrêa et al., 2025), and Juru (Junior et al., 2025) — have advanced the field, their application has largely focused on generative tasks. Although recent studies have employed NLP to detect biases in Brazilian courts (Benatti et al., 2024), the use of LLMs for structured information extraction remains underexplored compared to English-centric models like Lawma (Dominguez-Olmedo et al., 2025) and other similar contexts in the US (Frankenreiter and Livermore, 2025).

Efficient Fine-tuning: Processing the Brazilian judiciary's backlog of millions of lawsuits necessitates computationally efficient models. Techniques such as Q-LoRA (Quantized Low-Rank Adaptation) address the hardware constraint by allowing large models to be fine-tuned on consumer-grade hardware, freezing pre-trained weights while training only a small set of adapters. Complementing this architectural efficiency, recent findings from Lawma (Dominguez-Olmedo et al., 2025) demonstrate that fine-tuning base models on small (≈200 data points) but high-quality datasets can substantially increase accuracy in domain-specific tasks. This work integrates these strategies—parameter efficiency and data curation—to democratize access to jurimetrics, enabling high-performance analysis without prohibitive infrastructure costs.

3 Methodology

3.1 Data Acquisition and Corpus Construction

We focused our analysis on the Court of Justice of São Paulo (TJ-SP), the largest court in Brazil.[3]

Our dataset strategy involves two distinct components: **Training Set (Gold Standard):** Used as ground-truth to test and fine-tune models. We leveraged a high-quality legacy dataset from Lacerda e Silva (2021), consisting of 258 manually annotated judicial decisions from the same court. To maximize the volume of data available for the Q-LoRA (Dettmers et al., 2023) fine-tuning process, we prioritized the training split, randomly partitioning this corpus into 186 cases for training, 33 for validation, and 39 as a held-out test set. We note that this margin strictly quantifies subset-share representativeness; model extraction uncertainty is evaluated and reported separately via confidence intervals in Section 4.

Inference Set (Target Corpus): Using the `tjsp` R library (de Jesus Filho and Trecenti, 2020), we collected 12,981 first-instance sentences and then drew a stratified analytical sample of 454 documents split into a Control Group (early 2024, pre-decision) and a Treatment Group (early 2025, post-STF decision).[4]

3.2 Model Selection and Fine-Tuning

We employed the Unsloth framework (AI et al., 2025) to implement Q-LoRA fine-tuning. A crucial component of our pipeline was the integration of Pydantic (Colvin et al., 2026) for output validation, ensuring that generated JSONs strictly adhered to the defined schema during training. After an initial screening of over a dozen open-source models — evaluating them on reasoning capabilities and resource efficiency — we selected a commercial baseline (**GPT-4o-mini**, Team, 2024c) and two open-source candidates capable of running on consumer-grade hardware — **Llama 3.2 3B** (Team, 2024a) and **Phi-4 (14B)** (Team, 2024b).[5] We also observed promising zero-shot performance from the **Gemma 3** (Team, 2025) family; however, computational constraints prevented their fine-tuning in this experimental setup, reserving the adaptation of Gemma models as a suggestion for future work.

Selection of Phi-4 (14B): Comparative results demonstrated that **Phi-4 (14B)** offered the optimal balance. It demonstrated superior adaptation to the

[3] While this study focuses on drug trafficking cases (Law 11.343/06) in São Paulo, the proposed pipeline is domain-agnostic and can be adapted to other legal areas or tribunals provided that relevant data are collected

[4] At the document level, this sample corresponds to a 95% confidence level with an approximate margin of error of ± 4.52 percentage points for proportion estimates (conservative assumption $p = 0.5$, with finite population correction from $N = 12{,}981$)

[5] Ultra-lightweight models such as Llama 3.2 (1B) were discarded due to frequent syntax hallucinations during preliminary tests.

legal domain after fine-tuning— reducing syntax errors in comparison with Llama 3.2 (3B) — while remaining lightweight enough to run inference on a Google Colab Pro instance. Consequently, the fine-tuned Phi-4 (14B) was deployed to process two stratified samples from the corpus for the downstream statistical analysis.

4 Results and Discussion

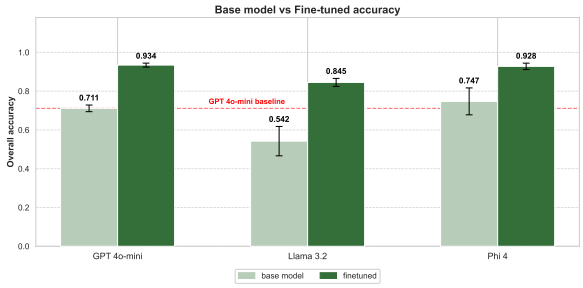

Figure 1: Overall extraction accuracy before and after fine-tuning. Error bars show 95% confidence intervals computed as $\pm 1.96 \cdot \sigma / \sqrt{n}$ over per-document accuracy ($n = 39$ test cases). The dashed red line marks the GPT-4o-mini base-model score as a zero-shot baseline reference.

Model	$F1_{bool}$	MCC	κ	MAE_{num}	$F1_{sent}$
GPT-4o-mini	.895 (.613)	.534 (.277)	.718 (.329)	27.4 (65.4)	.263 (.684)
Phi-4 (14B)	.826 (.679)	.430 (.343)	.551 (.372)	22.7 (6.9)	.960 (.396)
Llama 3.2 (3B)	.719 (.538)	.250 (.055)	.312 (.056)	35.1 (45.2)	.599 (.167)

Table 1: Complementary extraction metrics for fine-tuned (FT) and baseline (BL, in parentheses) models. Overall accuracy is reported in Figure 1. $F1_{bool}$ denotes macro-F1 over Boolean fields; $F1_{sent}$ denotes macro-F1 for the categorical verdict task (`sentenca`); lower MAE_{num} is better.

Figure 1 and Table 1 report performance metrics. To comprehensively evaluate the 47 fields mandated by our schema (which comprises 34 Boolean flags, 6 Numeric values, 5 Categorical labels, and 2 Text/NER fields), we compute both global and type-specific metrics. While global accuracy provides a baseline measure across all 1,833 extraction points, it can be artificially inflated by heavily imbalanced boolean fields (e.g., rare procedural flags that are *True* in fewer than 10% of cases). To mitigate this, we incorporate the Macro-averaged F1-Score, the Matthews Correlation Coefficient (MCC), and Cohen's κ. MCC jointly accounts for all four quadrants of the confusion matrix (where 0 indicates random guessing), while Cohen's κ rigorously discounts agreement attributable strictly to chance. The observable gap between high raw accuracy and moderate MCC/κ scores underscores the intrinsic difficulty of extracting rare legal phenomena, proving that strict accuracy alone is an insufficient metric for jurimetrics. For the 6 numerical fields (e.g., seized drug quantities in grams, sentence length in months), we report the Mean Absolute Error (MAE). Notably, the baseline Phi-4 (14B) achieved a deceptively low MAE (6.9) due to a zero-inflation artifact: lacking domain adaptation, the baseline often defaulted to zero or empty predictions, which coincidentally matched the ground truth for null drug seizures. In contrast, the fine-tuned models actively learned to extract real magnitudes, resulting in a strictly penalized but semantically meaningful numerical extraction.

4.1 Model Performance Evaluation

The fine-tuned Phi-4 (14B) model achieved an overall accuracy of **92.8%**, approaching the proprietary GPT-4o-mini FT (93.5%). The fine-tuned Llama 3.2 (3B) also demonstrated strong adaptation, reaching 85.4%, a massive improvement over its zero-shot baseline of 54.2%. To transparently report generalization uncertainty given the held-out test set, we computed 95% confidence intervals (CI) based on per-document accuracy, as illustrated in Figure 1 and Table 2. The non-overlapping confidence intervals between base and fine-tuned models suggest that specialized, resource-efficient smaller models can match or approach proprietary performance for structured legal extraction.

4.2 Jurimetric Impact Analysis

To assess the substantive legal impact of the Supreme Court's ruling (RE 635.659), we applied the fine-tuned Phi-4 (14B) model to process stratified samples drawn from the corpus of 12,981 sentences. The analysis was designed to detect statistical deviations between two distinct periods: the **Control Group** (sentences from early 2024, prior to the decision) and the **Treatment Group** (sentences from early 2025, post-implementation).

Because many Brazilian criminal sentences adjudicate multiple co-defendants simultaneously, the substantive unit of analysis is the individual defendant rather than the document. Our extraction pipeline therefore expands each parsed JSON array into defendant-level observations. The fully parsed Phi-4 (14B) FT dataset yielded 562 individual defendant records (286 in the 2024 control group and 276 in the 2025 treatment group), reflecting

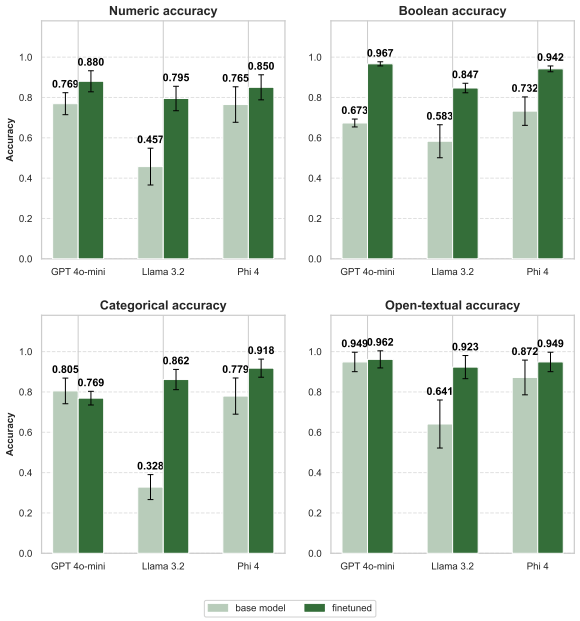

Figure 2: Per-task accuracy before and after fine-tuning [95% CI: $\pm 1.96 \cdot \sigma/\sqrt{n}$, $n = 39$]. Fine-tuning gains are consistent across most task types. Boolean fields show the largest gains (Llama-3.2 (3B): .58 $\rightarrow$ .85; Phi-4 (14B): .73 $\rightarrow$ .94). The only observed degradation was in GPT-4o-mini categorical (.81 $\rightarrow$ .77), driven by verdict classification ($F1_{sent}$: .684 $\rightarrow$ .263; Table 1).

reasonable variation in co-defendant counts across cases. This modest temporal imbalance (50.9% vs 49.1%) is unlikely to impact the validity of the non-parametric tests employed. Furthermore, while the fine-tuned Phi-4 (14B) successfully produced valid JSON arrays for all 454 documents, the Llama 3.2 (3B) FT exhibited a 2.2% structural failure rate (10 of 454 cases) when processing complex multi-defendant sentences. This empirical difference in strict schema adherence, alongside the broader accuracy metrics, supports our decision to adopt Phi-4 (14B) FT as the primary extraction engine for this jurimetric analysis.

The primary objective was to determine whether the distribution of key legal variables—specifically the frequency of drug trafficking convictions—exhibited significant shifts that would indicate effective adherence to the new binding precedent by first-instance judges.

We employed chi-square and Mann-Whitney U tests to assess changes across all variables, with particular attention to those most plausibly affected by the ruling, including conviction, mentions of marijuana, and sentencing outcomes. As shown in Figure 3, most comparisons were not statistically significant, although two tests yielded nominal p-values below 0.05. Because no correction for multiple comparisons was applied, these isolated results should be interpreted cautiously. Taken together, the results lead us to **fail to reject the null hypothesis** of no immediate measurable change in the main outcomes of interest. This pattern is consistent with short-run *institutional inertia*: despite the binding precedent aimed at decriminalizing possession for personal use, first-instance trafficking-conviction patterns remain stable in the immediate post-decision period.

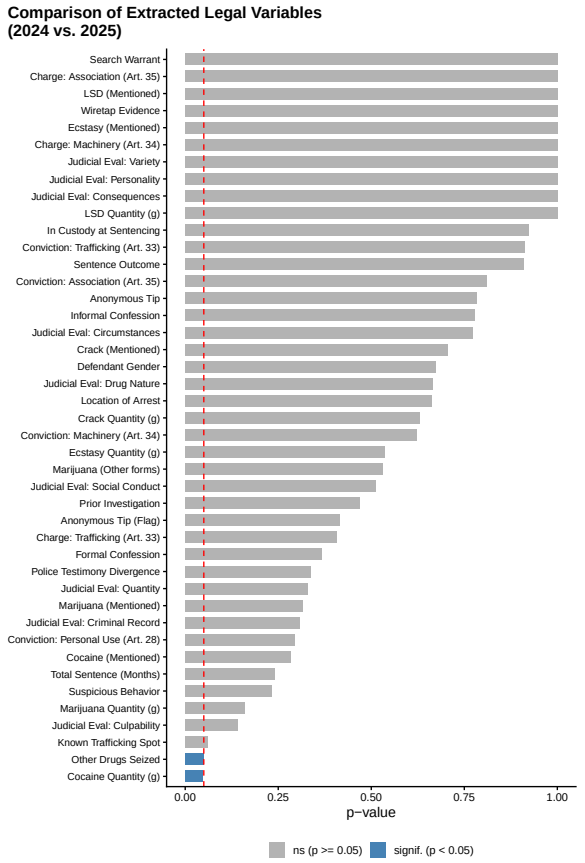

Figure 3: Statistical significance (p-values) of changes in legal variables (2024 vs 2025). Most variables show no significant change (grey bars), indicating stability in judicial behavior.

Substitution Effect: However, the automated analysis detected subtle adaptive behaviors in the micro-statistics of rulings. We observed nominal increases in two variables (uncorrected $p < 0.05$): *Other Drugs* and *Cocaine Quantity* (Figure 4). While these initial signals align with a potential "Substitution Effect"—where enforcement might shift toward "harder" drugs to maintain conviction rates as the legal threshold for marijuana increases—we note that these variables lose strict statistical significance when applying conserva-

tive multiple-testing corrections (e.g., Benjamini-Hochberg FDR).

Therefore, rather than definitive proof of institutional inertia, we present this substitution effect as an exploratory hypothesis. It highlights how LLM-powered jurimetric analysis can surface hidden micro-trends that warrant focused, causal follow-up studies with larger longitudinal datasets.

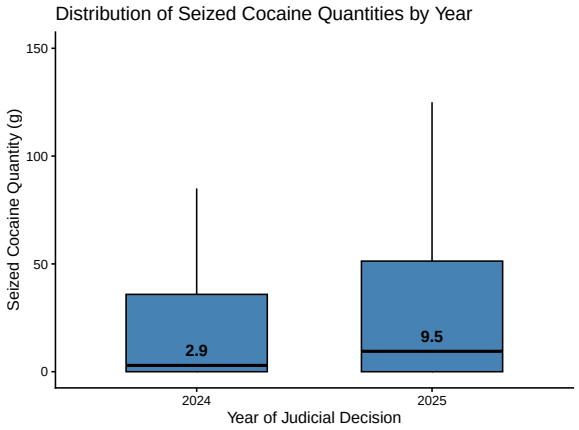

Figure 4: Boxplot of seized cocaine quantities (grams). The post-decision period (2025) shows a higher median and variance.

4.3 Methodological Implications

Our results empirically corroborate the hypothesis advanced by Dominguez-Olmedo et al. (2025) regarding the efficacy of "small-data" fine-tuning. We demonstrated that adaptation to the complex vernacular of Brazilian Courts was achieved with a dataset containing 258 documents —a volume feasible for individual annotation. By utilizing the Unsloth framework to optimize the Q-LoRA training process, we achieved this adaptation with high computational efficiency, eliminating the need for industrial-scale hardware. Furthermore, this workflow highlights the strategic value of repurposing legacy empirical research. By using a dataset from a prior study (Lacerda e Silva, 2021) as a gold standard, we advocate for a paradigm shift in legal scholarship: empirical legal studies should be strictly reproducible, as their curated data can now serve as the "ground truth" to bootstrap modern AI models. Crucially, this approach transforms unstructured text into reliable structured data, paving the way for a wide array of analytical and predictive tools—ranging from foundational statistics and advanced econometric frameworks to machine learning models—for monitoring judicial behavior at scale. This creates a scalable feedback loop: researchers can investigate the rich and complex dynamics of judicial decision-making—from auditing compliance with Supreme Court precedents to uncovering nuanced sentencing patterns—without relying on prohibitive manual labor.

5 Conclusion

This study establishes a dual contribution to the field of computational law. Methodologically, we validated that fine-tuning small, open-source models on repurposed legacy datasets yields useful performance. By achieving a 0.826 macro-F1 score alongside 92.8% overall accuracy in structured extraction via Q-LoRA, we provided evidence that high-performance NLP is not dependent on prohibitive budgets, thereby resolving critical barriers of cost and data privacy. From a jurimetric perspective, we proposed a scalable technique for legal analysis. The pipeline proved capable of converting vast amounts of unstructured text into rich, structured datasets, suggesting subtle phenomena—such as the 'substitution effect' in drug seizures—that traditional manual analysis would likely miss. Beyond merely auditing judicial compliance, this granular data extraction paves the way for advanced analytical and predictive methodologies. By unlocking these capabilities, the pipeline empowers researchers and policymakers to explore the complex realities of judicial behavior and monitor the real-world impact of legal precedents at an unprecedented scale.

6 Future work

Future work spans two fronts. First, expanding the annotated dataset will enable scaling Llama 3.2 (3B) fine-tuning and evaluating emerging models like Gemma 3. Second, future studies should leverage these structured datasets for advanced statistical modeling (e.g., clustering, regression) and transition to causal inference frameworks to rigorously test phenomena like the substitution effect.

Limitations

The main limitations are the relatively small test dataset, the geographic restriction to São Paulo (TJ-SP), and the short temporal scope (early 2025). Furthermore, our analysis evaluates macro-level shifts in general trafficking conviction rates, but it does not strictly isolate cases that perfectly match the STF's new decriminalization criteria (exclusive

marijuana possession under 40g). Finally, inferential statistics rely on model-extracted labels without propagating extraction uncertainty, meaning the observed substitution effect remains correlational and requires further causal investigation.

Ethical Considerations

Deploying LLMs for legal extraction requires careful ethical mitigation. First, models may inherit historical or alignment biases, shifting the role of human legal experts from manual annotators to critical validators. Second, to comply with the Brazilian General Data Protection Law (LGPD), the published dataset is robustly anonymized, masking Personal Identifiable Information (PII) such as defendant names and lawsuit numbers. Finally, because the models were fine-tuned on unredacted texts, their weights might retain sensitive data. To prevent misuse, they are released on Hugging Face via a gated access mechanism, requiring researchers to agree to strict privacy-preserving terms and explicitly prohibiting PII extraction.

Acknowledgments

The author would like to express profound gratitude to Prof. Lynn Kaack, PhD (Hertie School), and Prof. João Vitor Matos Gonçalves (University of São Paulo) for their supervision during the foundational research that originated this paper. The author is also deeply grateful to Prof. Mark Hallerberg, PhD, and Prof. Dr. José de Jesús Pérez Alcázar for their review and feedback as examiners of the original academic theses. Special thanks go to Marina Lacerda e Silva for generously making her annotated dataset available, and to the anonymous reviewers for their constructive comments.

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. Sabiá-3 technical report. *Preprint*, arXiv:2410.12049.

Unsloth AI, Daniel Han-Chen, and Michael Han-Chen. 2025. Unsloth. https://github.com/unslothai/unsloth.

Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-2: A New Generation of Portuguese Large Language Models. *arXiv preprint*. ArXiv:2403.09887 [cs].

Raysa Benatti, Fabiana Severi, Sandra Avila, and Esther Luna Colombini. 2024. Gender bias detection in court decisions: A Brazilian case study. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 746–763, New York, NY, USA. Association for Computing Machinery.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, and 1 others. 2026. Pydantic Validation.

Conselho Nacional de Justiça. 2025. Justiça em números 2025: Ano-base 2024. Technical report, Conselho Nacional de Justiça, Brasília.

Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2025. Tucano: Advancing neural text generation for portuguese. *Patterns*, 6(11).

Susana Henriques da Costa, Paulo Eduardo Alves da Silva, Marco Antonio da Costa Sabino, Débora Chaves Martines Fernandes, and Leonardo Augusto dos Santos Lusvarghi. 2011. A eficácia do sistema jurídico de prevenção e combate à improbidade administrativa. Publicação de relatório de pesquisa, Secretaria de Assuntos Legislativos do Ministério da Justiça, Brasília.

José de Jesus Filho and Julio Trecenti. 2020. tjsp: Coleta e organização de dados do tribunal de justiça de são paulo. https://github.com/jjesusfilho/tjsp.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna Gummadi, Moritz Hardt, and Michael Livermore. 2025. Lawma: The power of specialization for legal annotation. *Preprint*, arXiv:2407.16615.

Rodrigo Filippi Dornelles. 2025a. Enabling Jurimetrics: Deploying open-source Large Language Models for empirical insights into Brazilian courts. Master's thesis, Hertie School, Berlin, Germany. Supervisor: Prof. Lynn Kaack, PhD.

Rodrigo Filippi Dornelles. 2025b. Uso de inteligência artificial e jurimetria: uma aplicação ao caso das condenações por tráfico de drogas. Master's thesis, Universidade de São Paulo, São Paulo, Brazil. Supervisor: Prof. João Vitor Matos Gonçalves.

Jens Frankenreiter and Michael A. Livermore. 2020. Computational methods in legal analysis. *Annual Review of Law and Social Science*, 16. Virginia Public Law and Legal Theory Research Paper No. 2020-44, Virginia Law and Economics Research Paper No. 2020-09.

Jens Frankenreiter and Michael A. Livermore. 2025. Large language models in legal analysis. SSRN Electronic Journal. Virginia Public Law and Legal Theory Research Paper Forthcoming, Washington University in St. Louis Legal Studies Research Paper No. 25-09-01.

IPEA and Ministério da Justiça. 2023. Perfil do processado e produção de provas nas ações criminais por tráfico de drogas. Technical report, Instituto de Pesquisa Econômica Aplicada.

Roseval Malaquias Junior, Ramon Pires, Roseli Romero, and Rodrigo Nogueira. 2025. Juru: Legal Brazilian Large Language Model from reputable sources. *Preprint*, arXiv:2403.18140.

Marina Lacerda e Silva. 2021. Punindo as diferenças: Gênero, raça e geração no sentenciamento de tráfico de drogas na cidade de são paulo. Master's thesis, Universidade de Brasília.

Marcelo Guedes Nunes. 2016. Jurimetria: como a estatística pode reinventar o direito. *São Paulo: Revista dos Tribunais*, 2:30–30.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Supremo Tribunal Federal. 2024. Recurso Extraordinário 635.659. Brasília, DF. Relator: Min. Gilmar Mendes.

Gemma Team. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Llama Team. 2024a. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Microsoft Research Team. 2024b. Phi-4 Technical Report. *arXiv preprint*. ArXiv:2412.08905 [cs].

OpenAI Team. 2024c. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Appendix A: Model Benchmark

To justify the selection of the fine-tuned candidate models discussed in Section 4.1, we conducted a benchmark against 14 open-source and proprietary models. Table 2 details the full extraction metrics across all tested architectures.

Appendix B: Computational Efficiency and Environmental Impact

We measured energy consumption with NVML, polled by WandB at 15-second intervals, during inference over 454 judicial decisions on a single NVIDIA A100-PCIE-40GB GPU. Because the measurement includes idle power drawn by co-located GPUs on the same node, the reported values should be interpreted as a conservative upper bound. The main experiments were conducted in Berlin, Germany, between January and April 2025.[6]

Phi-4 (14B) FT achieves accuracy within 0.7 pp of GPT-4o-mini FT at **45% lower cost** (€0.77 vs. €1.39 per 1,000 cases); Llama 3.2 (3B) FT reduces cost by **71%** (€0.40/ 1000 cases) at 238 cases/ hour – the preferred configuration for large-scale institutional processing under privacy and data-sovereignty constraints.

Appendix C: Fine-Tuning Details

C.1 Environment

Both models were trained on a bare-metal server running Ubuntu 22.04, equipped with a 64-core CPU (128 logical threads) and 503 GB of RAM. Although the machine housed four NVIDIA A100-PCIE-40GB GPUs, only a single card was utilized per run (CUDA 12.2). The software stack included Python 3.10.12, Transformers 4.51.2, and Unsloth 2025.3.19. By applying 4-bit NF4 quantization via the `bitsandbytes` library and computing in `bfloat16` precision, both models fit entirely within the 40 GB VRAM footprint.

C.2 Q-LoRA Configuration and Hyperparameters

Parameter-efficient fine-tuning was applied via LoRA with rank r=16, α=16 ($\alpha/r = 1.0$), targeting all seven projection layers (`q/k/v/o/gate/up/down_proj`), no dropout, no bias. Both models were loaded in 4-bit NF4 quantization (double quantization enabled).

C.3 Training Outcomes

Phi-4 (14B) reached its best checkpoint in fewer optimization steps than Llama 3.2 (3B) and achieved a lower best `eval_loss` (0.0358 vs. 0.0878). This pattern is compatible with differences in model capacity and pretraining, and is consistent with Phi-4 (14B)'s higher downstream extraction accuracy (92.8% vs. 85.4%).

Appendix D: Description of features

[6] German non-household electricity tariff: €0.22/kWh (Eurostat, Q1 2024). GPT-4o-mini-FT cost was estimated with `tiktoken` using OpenAI Batch API pricing (€0.13/M input and €0.53/M output tokens, April 2025). Carbon emissions were estimated using a factor of 380 g CO_2eq/kWh (Agora Energiewende, Germany 2024).

#	Model	Accuracy	F1 Bool	F1 Sent	MCC	Kappa	MAE Num	Exact Match
1	GPT-4o-mini FT	0.935	0.895	0.263	0.534	0.718	27.4	0.0%
2	Phi-4 (14B) FT	0.928	0.826	0.960	0.430	0.551	22.7	5.1%
3	Llama 3.2 (3B) FT	0.854	0.719	0.599	0.250	0.312	35.1	0.0%
4	Gemma3 (27B	0.820	0.734	0.655	0.324	0.411	31.4	0.0%
5	GPT-4o	0.788	0.686	0.722	0.349	0.401	20.0	0.0%
6	o3-mini	0.764	0.698	0.684	0.370	0.422	28.0	0.0%
7	Gemma3 (12B)	0.763	0.668	0.572	0.259	0.311	370.8	0.0%
8	GPT-4.1	0.755	0.651	0.722	0.320	0.372	20.2	0.0%
9	Phi-4 (14B) baseline	0.747	0.679	0.396	0.343	0.372	6.9	0.0%
10	Llama 3.1 (8B)	0.735	0.668	0.315	0.294	0.332	23.4	0.0%
11	GPT-4o-mini baseline	0.711	0.613	0.684	0.277	0.329	65.4	0.0%
12	GPT-4.1-mini	0.697	0.632	0.217	0.279	0.335	19.3	0.0%
13	GPT-4.1-nano	0.659	0.558	0.307	0.224	0.247	37.8	0.0%
14	Llama 3.2 (3B) baseline	0.542	0.538	0.167	0.055	0.056	45.2	0.0%

Table 2: Comprehensive performance benchmark of 14 generative models on 39 held-out annotated sentences. **F1 Bool**: macro-averaged F1 over the 34 binary classification fields (drug presence flags, procedural flags, and judicial assessment indicators). **F1 Sent**: macro-averaged F1 for verdict classification (*sentença*), a 3–6 class categorical task. **MCC/Kappa**: global correlation and agreement metrics across all fields. **MAE Num**: mean absolute error on the 6 numeric fields (drug quantities in grams and sentence duration in months). **Exact Match**: percentage of documents where all 47 extracted values match the gold annotation exactly.

Table 3: Efficiency summary for 454 cases (single A100-PCIE-40GB).

Model	Time	s/case	kWh	Wh/case	Cost (€)	CO_2 (g/case)
Phi-4 (14B) FT	4h 08m	32.7	1.596	3.52	0.35	1.34
Llama 3.2 (3B) FT	1h 55m	15.1	0.814	1.79	0.18	0.68
GPT-4o-mini FT	n/a	n/a	n/a	n/a	0.63	n/a

Hyperparameter	Llama 3.2 (3B)	Phi-4 (14B)
Max scheduled epochs	10	10
Learning rate	2×10^{-5}	2×10^{-5}
LR scheduler	linear	linear
Warmup steps (script)	5	5
Effective warmup steps	5	46‡
Optimizer	AdamW-8bit	
Weight decay	0.01	
Max gradient norm	1.0	
Effective batch size	4 (1×4 accum.)	
Max sequence length	18,432 tokens	
Random seed	3407	
Eval / save every	10 steps	10 steps
Best-model criterion	`eval_loss`	
Chat template	`llama-3.2`	`phi-4`

Table 4: Training hyperparameters. ‡Phi-4 (14B): `SFTConfig` logged `warmup_ratio = 0.1`, overriding the script's 5 steps to $\max(5,\ 0.1 \times 460) = 46$.

Metric	Llama 3.2 (3B)	Phi-4 (14B)
Completed steps	460	330
Completed epochs	≈9.8	≈7.0†
Train loss (final step)	0.0183	0.0194
Train loss (epoch avg.)	0.1289	0.0408
Eval loss (best)	**0.0878**	**0.0358**
Final grad norm	0.137	0.077
Total FLOPs	2.17×10^{17}	7.74×10^{17}
Training time	58 min	118 min

Table 5: Training outcomes. †Phi-4 (14B)triggered early stopping at epoch 7; best checkpoint (`eval_loss` = 0.0358) was automatically restored.

Field name	English label	Description	Type	Task
juiz	Judge Name	Full name of the presiding judge, excluding titles or honorifics.	open_text	NER
sexo_juiz	Judge Gender	Gender of the presiding judge.	categorical	categorical
nome	Defendant Name	Full name of the defendant.	open_text	NER
local	Location of Arrest	Categorical location associated with the offense or arrest context.	categorical	multiclass_classif.
maconha	Marijuana (Mentioned)	Indicates whether marijuana was reported as seized in the case.	boolean	yesno_qa
maconha_g	Marijuana Quantity (g)	Reported quantity of marijuana, in grams.	numeric	numeric_qa
maconha_outras	Marijuana (Other forms)	Indicates whether other marijuana derivatives or forms, such as skunk or hashish, were mentioned.	boolean	yesno_qa
cocaina	Cocaine (Mentioned)	Indicates whether cocaine was reported as seized in the case.	boolean	yesno_qa
cocaina_g	Cocaine Quantity (g)	Reported quantity of cocaine, in grams.	numeric	numeric_qa
crack	Crack (Mentioned)	Indicates whether crack was reported as seized in the case.	boolean	yesno_qa
crack_g	Crack Quantity (g)	Reported quantity of crack, in grams.	numeric	numeric_qa
ecstasy	Ecstasy (Mentioned)	Indicates whether ecstasy was reported as seized in the case.	boolean	yesno_qa
ecstasy_g	Ecstasy Quantity (g)	Reported quantity of ecstasy, in grams.	numeric	numeric_qa
lsd	LSD (Mentioned)	Indicates whether LSD was reported as seized in the case.	boolean	yesno_qa
lsd_g	LSD Quantity (g)	Reported quantity of LSD, in grams.	numeric	numeric_qa
sentenca	Sentence Outcome	Final judicial outcome of the case.	categorical	multiclass_classif.
pena_base	Base Sentence	Base sentence as stated in the decision text.	categorical	NER
tot_pen	Total Sentence (Text)	Total sentence as stated in the decision text.	categorical	NER
tot_pen_meses	Total Sentence (Months)	Total sentence converted into months.	numeric	numeric_qa
outras_drogas	Other Drugs Seized	Indicates whether substances other than marijuana, cocaine, crack, ecstasy, and LSD were reported as seized.	boolean	yesno_qa
resultado_art_28	Conviction: Personal Use	Indicates whether the final decision applied Article 28 (personal drug use).	boolean	binary_classif.
resultado_art_33	Conviction: Trafficking	Indicates whether the final decision applied Article 33 (drug trafficking).	boolean	binary_classif.
resultado_art_34	Conviction: Machinery	Indicates whether the final decision applied Article 34.	boolean	binary_classif.
resultado_art_35	Conviction: Association	Indicates whether the final decision applied Article 35 (criminal association).	boolean	binary_classif.

Table 6: Evaluated extraction fields, English labels, descriptions, data types, and NLP task formulations (Part 1 of 2).

Field name	English label	Description	Type	Task
denuncia_art_33	Charge: Trafficking	Indicates whether the indictment included Article 33 (drug trafficking).	boolean	binary_classif.
denuncia_art_34	Charge: Machinery	Indicates whether the indictment included Article 34.	boolean	binary_classif.
denuncia_art_35	Charge: Association	Indicates whether the indictment included Article 35 (criminal association).	boolean	binary_classif.
flag_local_de_trafico	Known Trafficking Spot	Indicates whether the decision describes the location as associated with drug trafficking activity.	boolean	yesno_qa
flag_preso_no_momento_da_sentenca	In Custody at Sentencing	Indicates whether the defendant was in custody at the time of sentencing.	boolean	yesno_qa
flag_confissao_informal	Informal Confession	Indicates whether the text mentions an informal confession.	boolean	yesno_qa
flag_confissao	Formal Confession	Indicates whether the text mentions a formal confession, excluding informal admissions.	boolean	yesno_qa
flag_denuncia_anon	Anonymous Tip	Indicates whether the case mentions an anonymous tip.	boolean	yesno_qa
flag_denuncia	Formal Report	Indicates whether the case mentions a formal complaint or report, excluding anonymous tips.	boolean	yesno_qa
flag_atitude_suspeita	Suspicious Behavior	Indicates whether the defendant is described as exhibiting suspicious behavior.	boolean	yesno_qa
flag_divergencias_nos_relatos_dos_policiais	Police Test. Divergence	Indicates whether inconsistencies appear across police accounts.	boolean	yesno_qa
flag_investigacao	Prior Investigation	Indicates whether the case mentions prior investigation, police inquiry, or investigative steps.	boolean	yesno_qa
flag_interceptacao	Wiretap Evidence	Indicates whether interceptions, such as wiretaps or surveillance monitoring, are mentioned.	boolean	yesno_qa
flag_mandado	Search Warrant	Indicates whether the decision mentions a warrant being issued or executed.	boolean	yesno_qa
aval_antecedentes	Judicial Eval: Record	Indicates whether the decision explicitly evaluates the defendant's criminal record.	boolean	binary_classif.
aval_conduta	Judicial Eval: Conduct	Indicates whether the decision explicitly evaluates the defendant's conduct or social conduct.	boolean	binary_classif.
aval_personalidade	Judicial Eval: Personality	Indicates whether the decision explicitly evaluates the defendant's personality.	boolean	binary_classif.
aval_natureza	Judicial Eval: Drug Nature	Indicates whether the decision explicitly evaluates the nature of the offense or substances involved.	boolean	binary_classif.
aval_quantidade	Judicial Eval: Quantity	Indicates whether the decision explicitly evaluates the quantity of drugs or seized items.	boolean	binary_classif.
aval_variedade	Judicial Eval: Variety	Indicates whether the decision explicitly evaluates the variety of substances involved.	boolean	binary_classif.
aval_circunstancias	Judicial Eval: Circumst.	Indicates whether the decision explicitly evaluates the circumstances of the offense.	boolean	binary_classif.
aval_consequencias	Judicial Eval: Consequen.	Indicates whether the decision explicitly discusses the consequences of the offense.	boolean	binary_classif.
aval_culpabilidade	Judicial Eval: Culpability	Indicates whether the decision explicitly assesses culpability or blameworthiness.	boolean	binary_classif.

Table 7: Evaluated extraction fields, English labels, descriptions, data types, and NLP task formulations (Part 2 of 2).