Automated Audit of Judicial Metadata with Natural Language Processing: A Multi-Level Cascade Approach

Manuel Lucala Zengo[1], **Marlon Marcon**[2], **Fernando Ribeiro Trindade**[3]
Antonio Pires de Castro Júnior[4], **Willgnner Ferreira Santo**[5], **Manuel Findas**[6]

[1]Universidade Tecnológica Federal do Paraná (UTFPR) – Campus Cornélio Procópio
Cornélio Procópio – PR – Brazil

[2]Tribunal de Justiça do Estado de Goiás (TJGO)
Goiânia – GO – Brazil

[3]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

{sir.lucala,marlon.marcon}@gmail.com, manuelfinda@alunos.utfpr.edu.br,
fernandoribeirotrindade@gmail.com, apcastrojr@gmail.com,
willgnner.santos@discente.ufg.br

Abstract. *In the Brazilian judiciary, metadata such as case value, parties' names, and identification documents (CPF/CNPJ) are manually registered in the PROJUDI system alongside each filed petition, and clerks then manually verify this data against the petition—a process averaging 3 minutes per case. We propose a multi-level cascade combining high-precision regex, trained NER models, zero-shot extraction via GLiNER as a semantic judge, and a reverse search technique seeking expected metadata values within the petition text across multiple Brazilian formatting variants. Experiments on two real-case batches from the Court of Justice of Goiás (TJGO) show that, among validatable cases, extracted values agree with PROJUDI's registered metadata in 96% of cases in a noisier homologation batch (67 of 70 validatable, out of 185 total) and in 95.58% of a cleaner production batch (173 of 181 validatable, out of 182 total), at zero operational cost.*

1. Introduction

The digital transformation of the Brazilian Judicial Branch has progressed significantly in recent decades, driven by Law No. 11,419/2006, which regulated the electronic judicial process [Brasil 2006]. Systems such as PROJUDI (Digital Judicial Process), used by several state courts, digitize the procedural flow from filing to sentence; however, the quality of the registered metadata remains a relevant operational challenge.

When a lawyer files a new lawsuit in PROJUDI, two complementary actions are performed: (i) the manual registration of metadata in the system—including the case value, names of the active and passive poles, and identification documents (CPF/CNPJ)—and (ii) the upload of the initial petition in PDF format. Judicial clerks then manually verify whether the registered metadata corresponds to the actual content of the petition—a process that is time-consuming ($\sim$3 minutes per case), repetitive, inefficient at scale (thousands of cases per week), and prone to human error. At the Court of Justice of

the State of Goiás (TJGO), this operational bottleneck motivated the development of an automated solution.

In this work, we present an automated judicial metadata auditing system employing a multi-level cascade approach for information extraction and validation. Individually, the extraction techniques we combine are established in the NLP literature; as detailed in Section 2, the main contribution is the domain-specific, multi-dimensional validation rule that combines them into a single auditing decision, together with its deployment and evaluation in a real court production environment. More specifically:

1. A multi-dimensional validation that crosses five fields (case value, two procedural poles, and two identification documents) for a holistic compliance decision—to our knowledge not previously reported for judicial metadata auditing in the Brazilian judiciary;
2. An engineering integration of established extraction techniques into a cascade architecture—high-precision regular expressions, trained NER models with spaCy, and zero-shot extraction via GLiNER—combined for coverage rather than proposed here as new extraction methods in themselves;
3. A reverse search technique that complements traditional extraction, seeking the expected value in the text across multiple Brazilian monetary formatting variants;
4. Deployment and evaluation in a real production environment at TJGO, with zero operational cost—without dependency on paid APIs or cloud services.

The remainder of this article is organized as follows: Section 2 discusses related work; Section 3 details the proposed methodology; Section 4 presents the experiments and results; Section 5 discusses the findings and limitations; and Section 6 concludes the paper with future directions.

2. Related Work

The use of NLP techniques in the legal domain has received growing attention from the academic community, especially in the Brazilian context; this section reviews relevant work in legal NLP, named entity recognition, and automated document validation.

2.1. NLP Applied to the Brazilian Legal Domain

Named entity recognition (NER) in Brazilian legal texts was addressed by [de Araujo et al. 2018] with LeNER-Br, an annotated dataset for Portuguese legal NER (entities such as LEGISLATION, JURISPRUDENCE, PERSON) showing that domain-specific training significantly outperforms generic models. [Souza et al. 2020] presented BERTimbau, a pre-trained BERT model for Brazilian Portuguese widely used as a base for legal-domain NLP tasks, achieving state-of-the-art results on several Portuguese language-understanding tasks.

At the institutional level, the National Council of Justice (CNJ) has encouraged AI adoption in the Brazilian judiciary, notably via Resolution No. 332/2020, which sets ethical guidelines for its use [Conselho Nacional de Justiça 2020]. The Victor project at the Supreme Federal Court demonstrated the feasibility of automatic case classification [**?**]. Within state courts, the Berna system (Recursive Electronic Search using Natural

Language) [**?**] was implemented at TJGO to semantically group and triage similar initial petitions against predatory and mass litigation, demonstrating institutional NLP for case triaging and supporting complementary metadata-cleansing tools at the same court. Additionally, the INACIA system (AI-Assisted Legal Instruction) [**?**], developed for the Brazilian Federal Court of Accounts (TCU), applies language models to admissibility validation and automated fact extraction from initial pleadings, corroborating the relevance of human-in-the-loop decision-support systems in this space.

2.2. Information Extraction in Legal Documents

Information extraction from legal documents presents unique challenges, such as specialized vocabulary, complex textual structures, and formatting variations [Dale 2019]. Hybrid approaches combining linguistic rules with machine learning have proven effective in this domain [Chiticariu et al. 2013]. The spaCy library has consolidated itself as one of the main tools for industrial NLP, offering pre-trained Portuguese models and custom NER training support [Honnibal et al. 2020], with an architecture suited to efficient large-scale pipelines.

In this context, hybrid pipelines combining multiple extraction paradigms are recommended for a better accuracy/performance balance. As demonstrated by [**?**] in sensitive data discovery for Portuguese, strict-format identifiers (document numbers, verification codes) are best validated via rule heuristics and regex with checksum validation, while supervised NER is reserved for nominal and contextually ambiguous classes—substantially reducing the demand for neural processing. Furthermore, to handle the absence of rigid templates in heterogeneous documents, contextual relevance scoring is a promising alternative: in template-free structured data extraction [**?**], candidate entities are selected via a consolidated score integrating positional (distance to label), logical (syntactic compliance), and semantic (BERT cosine similarity) components—exhibiting high resilience for local processing of judicial documents.

2.3. Zero-Shot Models for NER

Recently, zero-shot NER models have emerged as promising alternatives that bypass the need for domain-specific annotated data. GLiNER, proposed by [Zaratiana et al. 2023], is a lightweight, generalist NER model enabling arbitrary entity extraction without fine-tuning; its ability to operate with textual labels at inference time makes it particularly suitable as a fallback mechanism in cascade pipelines.

2.4. Positioning of this Work

We want to be precise about where the novelty of this work lies. Individually, the extraction techniques we use—regex cascades, trained NER, zero-shot NER—are established in the literature (e.g., [**?**] already combines rule-based extraction with supervised NER similarly). What distinguishes this work (Table 1) is that it targets **validation** of already-registered metadata against the petition text via a multi-dimensional business rule jointly checking five fields—not extraction of an unknown value in isolation. To our knowledge, this combination, deployed in real production at zero cost, has not previously been reported for the Brazilian judiciary.

Table 1. Positioning relative to related institutional NLP systems, as reported in the cited sources.

System	Task	Technique / cost, as reported
Berna [?]	Semantic grouping/triage of similar petitions	NLP-based semantic search; models/cost not detailed
INACIA [?]	Admissibility validation, fact extraction	LLM-based; cost-incurring, cloud-dependent
This work	Multi-field compliance auditing of metadata vs. petition text	Regex + trained NER + zero-shot GLiNER + reverse search; fully local, zero cost

3. Methodology

This section details the proposed architecture, the processing pipeline, and the techniques employed for automated judicial metadata auditing.

3.1. Architecture Overview

The system consists of five sequential stages: (1) intelligent data loading; (2) cleaning and normalization; (3) multi-level extraction; (4) multi-dimensional cross-validation; and (5) classification and reporting.

The system input is a CSV file containing, for each judicial case, the metadata registered in PROJUDI (case value, names of the active and passive poles, CPF/CNPJ) and the full text of the initial petition (`inteiro_teor`), previously extracted from the PDF by Apache Tika.

3.2. Preprocessing and Cleaning

Preprocessing addresses three domain-specific challenges. **Robust parsing of malformed CSVs:** the CSV files exported by the TJGO system frequently contain fields with internal newlines, unescaped double quotes, and fields exceeding 128 KB (full petitions), so the system first attempts a standard load via `pandas` and, if the proportion of null values exceeds 50%, automatically falls back to a robust regex-based parser that recovers the malformed records. **Encoding correction:** texts extracted via OCR frequently present *Mojibake* (UTF-8 interpreted as Latin-1), which a dictionary of substitutions and heuristic re-decoding attempts to restore. **Textual normalization:** markers inserted by Apache Tika (`>>>>>inicio<<<<<`, `#####fim#####`) are removed, whitespace is normalized, and records with insufficient text are filtered out.

3.3. Multi-Level Case Value Extraction

Case value extraction employs a cascaded strategy with three levels of increasing sophistication, escalating from one level to the next on failure or deviation: cascaded regular expressions, then trained NER, then GLiNER as a semantic judge.

Level 1 — Cascaded Regular Expressions. The first level utilizes regular expressions organized by degree of confidence:

- **High confidence (∼99%):** Patterns capturing canonical formulations such as "Dá-se à causa o valor de R$ X", "Atribui-se à causa o valor de R$ X", "Valor da causa: R$ X".
- **Medium confidence (∼90%):** Less specific variations of monetary patterns in legal contexts.
- **Values in words (∼95%):** A dedicated parser that converts text numbers written in words to decimal values (e.g., "vinte e três mil reais" → 23,000.00).

The regex module also incorporates an OCR corrector that identifies and repairs common character substitutions in digitized texts (`O`→`0`, `S`→`5`, `I`→`1` in numerical contexts).

Level 2 — Trained spaCy NER. When Level 1 fails, the system resorts to a Named Entity Recognition model trained on spaCy's `pt_core_news_lg` base. The model was customized to detect the `VALOR_CAUSA` entity in legal petitions, utilizing 92 auto-annotated samples and 30 training iterations.

Level 3 — GLiNER as Semantic Judge. As a final fallback mechanism, the system employs the GLiNER model (`gliner_small-v2.1`) [Zaratiana et al. 2023] for zero-shot entity extraction. The model is queried with the labels "valor da causa" and "valor da ação", using a threshold of 0.3. GLiNER operates as a semantic judge: when triggered (difference $> 1\%$ between the regex extracted value and the metadata), it seeks to confirm or refute the candidate value.

The loading of GLiNER is lazy and thread-safe, avoiding memory consumption when not needed and allowing concurrent operation in a production environment.

3.4. Reverse Search — The Complementary Approach

In addition to direct extraction, the system implements a reverse search technique that operates in the opposite direction: instead of extracting values from the text and comparing them with metadata, it searches for the expected values (registered metadata) within the petition text.

For the case value, the system generates multiple textual representations of the registered value (e.g., for R$ 1,234.56: "1.234,56", "1234,56", "1234.56") and checks if any appear in the text.

For the procedural poles, the registered names are normalized (removal of accents, conversion to lowercase, removal of punctuation) and searched as substrings in the normalized text of the petition. The system supports joint litigants, where multiple names are separated by "#"; finding any of the names is sufficient to confirm compliance.

For the CPF/CNPJ, the documents are searched as sequences of digits in the text, with automatic restoration of leading zeros lost during spreadsheet exports.

Because the search string is derived from the registered metadata value itself, a match confirms that the corresponding sequence of characters occurs somewhere in the petition text, but it does not, by itself, guarantee that this occurrence is the genuine reference to the case value, party name, or document—rather than, for instance, an unrelated number that coincides with the same digit sequence. We discuss the implications of this limitation for the evaluation of reverse search in Section 5.2.

3.5. Risk Classification

The system classifies each case into one of two categories, according to the business rule defined cooperatively with TJGO clerks:

- **DIRECTLY APPROVED:** case value percentage difference $< 1\%$, AND compliant poles, AND compliant documents, AND no critical errors, AND at least one positive signal found.
- **WITH DEVIATION:** any other case—including deviations in any field, missing data, or processing errors.

For each of the five validated fields (value, active pole, passive pole, active CPF/CNPJ, passive CPF/CNPJ), the system assigns an individual status among five states: (1) **COMPLIANT** (registered data present and equivalent in the text); (2) **DIVERGENT** (registered data present, but different from text); (3) **NOT FOUND** (registered data not located in text or absent in registry); (4) **NOT APPLICABLE** (field not pertinent to the procedural class); and (5) **INVALID REGISTRY DATA** (registered field contains corrupted/leaked CSV text).

4. Experiments and Results

4.1. Dataset and Experimental Setup

To evaluate the performance and robustness of the developed solution, the system was validated in two distinct scenarios based on real TJGO production data:

- **Batch 1 (Homologation Environment):** Composed of 185 judicial processes. This batch is characterized by a high index of registry inconsistencies arising from formatting errors in the original CSV (*upstream*), serving as a stress test for filtering corrupted data. The distribution includes 70 cases with a registered value and validatable class, 14 interlocutory appeals (without case value), and 110 cases with at least one corrupted metadata field.
- **Batch 2 (UPJ Juizados da Fazenda Pública Production):** Composed of 182 judicial processes distributed across the 1st, 2nd, 3rd, and 4th Juizados da Fazenda Pública of Goiânia (members of the 1st Permanent Justice 4.0 Center). The collection covered the period from 25/05/2026 to 29/05/2026, representing a scenario with integral registry metadata.

The system was deployed in a Docker container running Python 3.10, spaCy 3.7.5 (`pt_core_news_sm`), GLiNER 0.1.11 (`gliner_small-v2.1`), and Gunicorn as the WSGI server.

4.2. General Results

Table 2 presents the comparative results of the binary classification obtained by the system across both validation batches, together with the metadata–petition agreement rate in case value extraction among validatable cases only (those with a registered value and whose procedural class supports case value attribution). We report the latter as an *agreement rate* rather than an accuracy rate—a case counts as agreement whenever the percentage difference is below 1%, which is not the same as an independently verified correct value (Section 5.2).

Table 2. General classification and case-value agreement results in the validation batches.

Classification	Batch 1 (Homologation)		Batch 2 (Fazenda Pública)	
	Quantity	Percentage	Quantity	Percentage
Directly Approved	44	24%	111	60.99%
With Deviation	141	76%	71	39.01%
Total	**185**	**100%**	**182**	**100%**
Case value, validatable cases only:				
Difference $< 1\%$ (agreement)	67	96%	173	95.58%
Difference $\geq 1\%$ (divergence)	3	4%	8	4.42%
Total validatable	**70**	**100%**	**181**	**100%**

The significant variation in the directly approved rate without human review (from 24% in Batch 1 to 60.99% in Batch 2) stems directly from database integrity: in Batch 2, the absence of systemic corruption in the metadata export allowed the majority of processed cases to be archived automatically, confirming the operational viability of the system, while cases routed to "With Deviation" correctly retained pole or document discrepancies. Analyzing the root cause of Batch 1's 141 deviations, the large majority (109 cases, 77%) originate outside the system—pole/CPF fields contaminated by unescaped double quotes in the source *upstream* CSV—while 16 (11%) are legitimate deviations and 16 (11%) involve institutional defendants without a CNPJ; the system detects and labels the corrupted-source cases as "INVALID REGISTRY DATA", making the root cause visible to the auditor.

4.3. Metadata–Petition Agreement Rate in Case Value Extraction

The agreement rates of 96% (Batch 1) and 95.58% (Batch 2) indicate consistency of the multi-level hybrid strategy across batches with different case classes and financial distributions. Given the modest validatable-case counts (70 and 181), we report Wilson 95% confidence intervals: [88.1%, 98.5%] for Batch 1 and [91.5%, 97.7%] for Batch 2—overlapping, consistent with (though not proof of) the same underlying rate. This is measured against the registered metadata itself, not an independently verified ground truth (Section 5.2). In cases of divergence ($\geq 1\%$), inspection pointed to unreadable text blocks or atypical formatting as the likely cause, correctly flagged for human audit; we cannot, however, rule out that some instead reflect an incorrect registered value rather than an extraction failure.

4.4. Status Distribution by Field

Table 3 presents the comparative distribution of individual span statuses in the validation batches.

In Batch 1, nearly 30% of individual spans were labeled as "INVALID REGISTRY DATA" due to structural corruption in the original input file. In Batch 2, database integrity resulted in zero occurrences of this failure, raising the direct compliance rate (*COMPLIANT*) to 88.35%—again, agreement with registered metadata rather than independently verified correctness (Section 5.2).

Table 3. Comparative distribution of individual statuses in validation spans.

Span Status	Batch 1 (Homol.) Qty.	Batch 1 (Homol.) Perc.	Batch 2 (Fazenda P.) Qty.	Batch 2 (Fazenda P.) Perc.
COMPLIANT	350	37.80%	804	88.35%
DIVERGENT / NOT FOUND	288	31.13%	103	11.32%
INVALID REGISTRY DATA	273	29.51%	0	0.00%
NOT APPLICABLE / CRITICAL ERROR	14	1.51%	3	0.33%
Total Spans	**925**	**100%**	**910**	**100%**

4.5. Per-Field Breakdown for Batch 2

Table 3 aggregates COMPLIANT/DIVERGENT status across all five field types. To make the per-field compliance rates cited later traceable, Table 4 breaks Batch 2 down by field, computed from that run's per-case audit log (182 cases), with Wilson 95% confidence intervals. We lack the equivalent log for Batch 1, so this breakdown covers Batch 2 only; archiving per-case logs for future batches is future work (Section 6).

Table 4. Per-field compliance in Batch 2 ($n = 182$), with Wilson 95% confidence intervals.

Field	Compliant/Total	Rate	95% CI
Active pole (name match)	164/182	90.11%	[84.9%, 93.7%]
Passive pole (name match)	153/182	84.07%	[78.1%, 88.7%]
Active CPF/CNPJ	162/182	89.01%	[83.6%, 92.8%]
Passive CPF/CNPJ	152/182	83.52%	[77.4%, 88.2%]

4.6. Contribution of Each Extraction Method

The (approximate) contribution of each method in the case value extraction pipeline is: high-confidence regex, Level 1a ($\sim$60% of cases, $\sim$99% confidence); medium-confidence regex, Level 1b ($\sim$10%, $\sim$90%); trained spaCy NER, Level 2 ($\sim$15%, $\sim$94%); reverse search, complementary ($\sim$10%, $\sim$95%); and not found ($\sim$5%). Level 1 (regex) alone resolves the majority of cases ($\sim$70%), confirming that well-defined textual patterns are highly effective in the legal domain, while reverse search acts as a safety net for cases whose textual formulation deviates from expected patterns but whose numerical value is still present in the text.

This breakdown omits GLiNER (Level 3): batch-level activation logs for the semantic-judge layer were only retained for Batch 2, where it triggered in none of the 182 cases (Section 4.8). We independently verified this by re-running Level 1 (regex) and reverse search alone—NER (Level 2) and GLiNER (Level 3) both disabled—against the raw Batch 2 petitions: it exactly reproduces the published classification and per-field numbers (111/71 in Table 2; the four rates in Table 4), confirming empirically that Levels 2–3 add nothing beyond Level 1 and reverse search there. Batch 1—the noisier batch, where a semantic-judge fallback should matter most—lacks the per-case log for the same check, so the three-level cascade's design remains undemonstrated on the batch that would test it; retaining per-level logs for a full ablation on both batches is future work (Section 6).

4.7. Performance and Cost

The average processing time is 2–3 minutes for 1,000 records, compared to $\sim$3 minutes per case in manual verification. The operational cost is R$ 0.00—the system operates entirely locally, without dependency on paid APIs, commercial models, or cloud services.

In an earlier, internal test with $\sim$1,000 cases, we observed an automatic-approval rate of $\sim$76%. Unlike Batches 1 and 2, this test does not have a preserved batch description, extraction-method breakdown, or root-cause analysis, so we report it only as a preliminary, undocumented observation, not as a validated scalability result or as evidence of approval accuracy. A controlled scalability study with the same documentation as Batches 1–2 is listed as future work (Section 6).

4.8. Comparative Benchmarks with the Literature

To situate this work relative to related literature, we juxtapose our rates below with reported baselines for Portuguese text processing. This is a juxtaposition of differently defined metrics, not a like-for-like comparison: the F1-Scores below were computed against annotated ground truth, whereas ours measure agreement with metadata already registered in PROJUDI (Sections 4.3, 5.2).

- **Template-Free Extraction:** [**?**] obtained an F1-Score of 0.71 with local spaCy fine-tuning and 96% precision in label matching—a similar numerical range to our case value agreement rate (96% Batch 1, 95.58% Batch 2), though not directly comparable since ours is not evaluated against annotated ground truth.
- **Sensitive Data and Identifier Extraction:** [**?**] reported an F1-Score of 83.01% for sensitive data discovery in Portuguese using Bi-LSTM. Our document compliance rates (Batch 2, Table 4) of 89.01% (active CPF/CNPJ) and 83.52% (passive) do not require heavy neural models, subject to the same caveat.
- **AI Infrastructure Cost:** Unlike systems relying on commercial LLMs or high-cost compute [**?**], GLiNER activation was 0% in valid Batch 2 cases, indicating a local, zero-cost architecture is practical at this scale.

5. Discussion

5.1. Technical Contributions

The main technical contribution of this work is demonstrating that a pragmatic cascaded approach—combining simple techniques (regex) with more sophisticated models (NER, GLiNER)—can achieve high metadata–petition agreement rates in the legal domain without incurring generative AI infrastructure costs.

The reverse search technique (Section 3.4) deserves particular highlight: unlike the traditional *extract-and-compare* paradigm, it operates in the opposite direction, *verifying whether the expected value exists in the text*—especially effective for validation (confirming a known value) rather than discovery (finding an unknown one).

Using GLiNER as a zero-shot semantic judge is also a relevant contribution: unlike conventional use of zero-shot models as primary extractors, our system employs them as *arbitrators of last resort*, triggered only when more efficient methods fail, optimizing computational resource usage without sacrificing coverage.

5.2. What the Reported Rates Do and Do Not Measure

The 95.58%/96% figures (Section 4.3) and the compliance rates throughout Section 4 should not be read as accuracy against an independently verified ground truth. They measure agreement between the extracted value and the value already registered as metadata in PROJUDI, under a 1% tolerance. A disagreement can mean an extraction error, an incorrect registered value, an ambiguous petition, or an OCR error; conversely, coinciding values could still both be wrong if the same incorrect amount was captured by both clerk and extractor. Without independent annotation, these cannot be distinguished from the data reported here.

A rigorous accuracy evaluation would require clerks or trained annotators to independently determine, for a petition sample, the correct case value, parties, CPF/CNPJ, PROJUDI compliance, and whether the case should have been approved directly—then compare the system's output against this gold standard, which is not yet available to us. We treat this as the most important future-work item (Section 6), and use "agreement" rather than "accuracy" throughout wherever the comparison is against registered metadata rather than a verified ground truth.

Relatedly, Table 2 reports the distribution of the system's own output classes, not a performance measurement: we do not report true/false approvals, precision/recall for With Deviation, specificity, or false-approval rate against a gold standard. False approval—incorrect metadata accepted without review—is the most operationally costly error, and this study gives no direct evidence of its frequency; estimating it is part of the same future gold-standard evaluation.

Finally, because reverse search (Section 3.4) queries the petition text using the registered value itself, judging it correct whenever it reproduces that value risks circularity: it could be locating an unrelated occurrence of the same digit sequence rather than a genuine correspondence. Distinguishing these cases likewise needs a human-verified reference independent of the search query.

5.3. Limitations

Beyond the evaluation gap discussed above, the study presents further limitations that should be considered:

1. **Input data quality:** Upstream CSV corruption significantly impacts the results (77% of deviations in Batch 1). Resolving this issue is external to the system.
2. **Dataset size:** The validation batches, while containing real cases, represent a limited sample. Evaluations with larger volumes are required.
3. **NER model with few samples, no held-out split:** The spaCy model was trained with only 92 auto-annotated samples over 30 iterations, all used for training, with no separate development or test split; so its coverage and overfitting risk cannot currently be assessed independently of the end-to-end pipeline results. We also did not measure inter-annotator agreement for the auto-annotation itself, nor run an ablation isolating NER's marginal contribution from reverse search—all listed as future work.
4. **Name spelling:** The system treats spelling variations as deviations (e.g., "Zacarias" $\neq$ "Zacharias"), a deliberate prioritization of security over convenience in an audit context.

5. **Code evolution since reported experiments:** Business rules have continued to evolve after these batches were processed (e.g., a later change excludes Agravo de Instrumento cases from case-value validation). Re-running the current code against Batch 2 still reproduces the published classification and per-field numbers (Section 4.5), but this may not hold for Batch 1 or future batches; we recommend tagging the code revision used per reported batch going forward.
6. **Portability and code availability:** The system was designed and evaluated only against PROJUDI's CSV export format at TJGO; adapting it to another court's export format is expected to require nontrivial, unquantified engineering effort. The source code is intended to be released, but no repository or timeline is available yet.

5.4. Comparison with Commercial Solutions

Unlike LLM-based solutions (e.g., GPT-4, Claude), which incur token costs and depend on cloud connectivity, our system operates entirely locally—a substantial cost difference for a court processing thousands of cases per week. Judicial data also frequently contains sensitive information (CPFs, party names) whose transmission to external servers raises compliance questions under the General Data Protection Law (LGPD).

6. Conclusion and Future Work

This work presented an automated judicial metadata auditing system employing a multi-level cascade approach for extracting and validating information in initial petitions. The system combines high-precision regular expressions, trained NER models, zero-shot extraction via GLiNER, and reverse search to validate five metadata fields simultaneously.

Experimental results with real TJGO cases demonstrate: (i) an agreement rate of 95.58% between extracted and registered case values among validatable cases, which we distinguish from validated accuracy in Section 5.2; (ii) processing capacity of $\sim$1,000 cases in 2–3 minutes, compared to $\sim$3 minutes per case in manual verification; and (iii) zero operational cost. Root cause analysis reveals that the majority of deviations originate from input data corruption, not system failures.

As future work, we intend to: (1) conduct a gold-standard evaluation with judicial clerks or trained annotators independently labeling case values, parties, and CPF/CNPJ for a sample of petitions, enabling true accuracy, precision, recall, and false-approval rate for both extraction and the final auditing decision (Section 5.2); (2) expand the NER training set to 500+ annotated samples with a proper train/development/test split and reported inter-annotator agreement, and run an ablation isolating NER's marginal contribution from reverse search and GLiNER on a noisy batch comparable to Batch 1 (Section 4.5), retaining per-level cascade logs for this purpose; (3) run a controlled scalability study, with the same batch documentation as Batches 1–2, to substantiate the preliminary $\sim$1,000-case throughput observation (Section 4.7); (4) evaluate portability to Brazilian courts using case-management systems other than PROJUDI, ideally via direct API integration rather than CSV export; and (5) improve pole validation with fuzzy/token-overlap matching and an institutional-defendant whitelist (e.g., INSS, Caixa Econômica, União), and evaluate large language models as an additional cascade layer for ambiguous cases.

The system's source code is in the process of being made available to the academic community.

References

Brasil (2006). Lei nº 11.419, de 19 de dezembro de 2006. Dispõe sobre a informatização do processo judicial. Diário Oficial da União, Brasília, DF. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11419.htm.

Chiticariu, L., Li, Y., and Reiss, F. R. (2013). Rule-based information extraction is dead! Long live rule-based information extraction systems! *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 827–832.

Conselho Nacional de Justiça (2020). Resolução nº 332, de 21 de agosto de 2020. Dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário. Diário da Justiça Eletrônico, Brasília. Disponível em: https://atos.cnj.jus.br/atos/detalhar/3429.

Dale, R. (2019). Law and word order: NLP in legal tech. *Natural Language Engineering*, 25(1):211–217.

de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 313–323. Springer.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. In *Zenodo*. DOI: 10.5281/zenodo.1212303.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. Springer.

Zaratiana, U., Ngouda, N., Ayach, A., Tomeh, N., and Charnois, T. (2023). GLiNER: Generalist model for named entity recognition using bidirectional transformer. *arXiv preprint arXiv:2311.08526*.