

A Hybrid Framework for Precedent Reference Extraction

Arthur Elwing Torres^{†*1}, Johnny Moreira^{*2}, Daniela Vianna², Marina Ramalhete Masid²,
Celio Larcher², Edleno de Moura^{1,2}, Altigran da Silva¹

¹Universidade Federal do Amazonas, ²Jusbrasil

{arthur.torres, alti, edleno}@icomp.ufam.edu.br

{johnny.moreira, daniela.vianna,
marina.ramalhete, celio.larcher}@jusbrasil.com.br

[†]Corresponding author. ^{*}Equal contribution.

Abstract

Legal documents frequently contain heterogeneous references to judicial precedents, making automated Precedent Reference Extraction (PRE) challenging for legal AI systems. Existing approaches rely heavily on LLM-based pipelines, which are often effective but also incur recurring prompting costs in high-volume production workloads and struggle with tracing extracted entities back to their character offsets in the source text. In light of these issues, we present *Notarius*, a domain-specific and production-ready hybrid framework that decomposes PRE into two stages: Overlapped NER Feature Extraction (ONFE), which extracts fine-grained precedent features, and Precedent Feature Aggregation (PFA), which reconstructs complete references for downstream applications. Our implementation integrates the Unified NER method W²NER with domain-specific heuristics. We evaluate *Notarius* at both feature and reference levels on a collection of judicial decisions and compare it in production against an LLM-based pipeline. *Notarius* achieves higher extraction quality, eliminates recurring LLM API costs for PRE, and provides span-level offset traceability. The system currently powers precedent extraction on Jusbrasil, demonstrating that discriminative, domain-specialized NER architectures remain valuable for cost-effective, high-precision legal information extraction in the LLM era.

1 Introduction

In modern legal systems, precedents consist of collections of past judicial decisions, interpretations, and rulings issued by courts. They play a central role in promoting legal stability and predictability, enabling judges and legal practitioners to resolve new cases by drawing on established judicial understanding. In this work, we address the problem of accurately extracting precedent references, a task we refer to as *Precedent Reference Extraction*

(PRE), in legal documents from the Brazilian civil-law system, which constitutes the primary domain served by Jusbrasil¹.

Effective PRE is essential for downstream applications such as legal search, precedent resolution, and automated document analysis. While PRE can be framed as a Named Entity Recognition (NER) task, directly extracting precedent references is challenging due to the absence of uniform citation standards and the diversity of judicial acts that may constitute a precedent, including court rulings, settled case law, and binding summaries (Correia et al., 2022; Mentzingen et al., 2025). As a result, precedent citation strings, which act as the primary reference unit in legal documents by encompassing multiple features necessary for uniquely identifying a precedent, are written in a wide variety of reference formats, as illustrated in Figure 1.

Another challenge in PRE is information fragmentation, which leads to discontinuous NER (Dai et al., 2020; Yan et al., 2021; Li et al., 2022), where a single reference is spread across non-contiguous spans. As seen in Figure 1, the first reference contains all features required for unique precedent identification within (overlapping) the citation string, whereas the second includes only the *CLASS* and *NUMBER*, with the *COURT*, *VENUE*, and *STATE* appearing earlier in the sentence. Traditional flat NER models cannot resolve overlapped or discontinuous entities, and regex-based approaches become overly complex to design and maintain for such structures. Modern Unified NER methods can detect these types of entities, but they remain impractical for large-scale PRE due to the high cost and complexity of annotating discontinuous

¹Jusbrasil is a major legal-technology (legaltech) platform, reporting more than 30 million users and reaching approximately 80% of the country’s lawyers. Although built around judicial and legal information, the majority of its traffic originates from ordinary citizens outside the legal profession, positioning it as a mass-market gateway to legal content rather than a strictly professional tool.

Reference is made to the legal understanding established in **Precedent 123 of the Supreme Federal Court**,

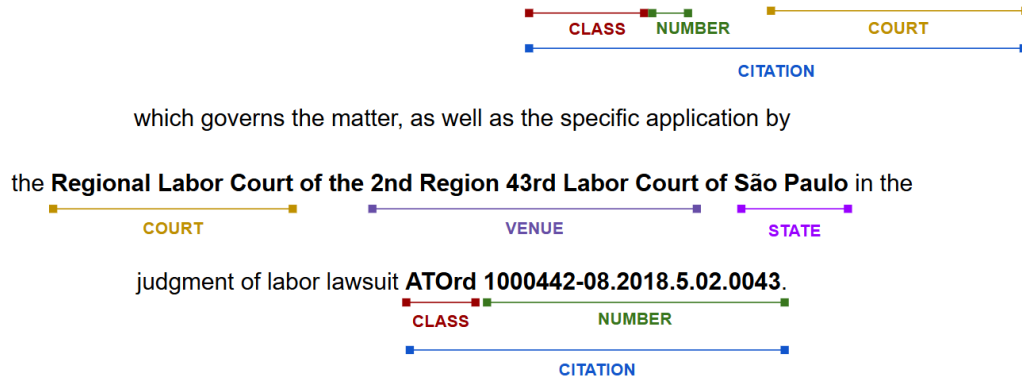


Figure 1: Snippet of a translated Brazilian judicial decision illustrating two precedent references in distinct formats along with their respective features.

spans. While LLM-based NER was initially explored within Jusbrasil and proved effective at capturing these overlapped and fragmented entities, empirical long-term deployment revealed some structural and financial bottlenecks. First, because company applications run continuously across high volumes of documents, the variable token costs of generative prompting create an unsustainable financial overhead compared to local discriminative models (Zschache and Hartwig, 2026; Alaswad et al., 2026). Second, reliably recovering exact character-level offsets from generative LLM outputs remains challenging (Wang et al., 2025; Zhan et al., 2026). This limitation was also observed during the deployment of our previous LLM-based pipeline, where generated extractions could not consistently be mapped to their exact spans in the source document. Such span-level traceability is a functional requirement within the company ecosystem: downstream applications rely on exact starting and ending character offsets for entity linking, automatic text highlighting, and dynamic hyperlinking.

In this paper, we propose a practical and scalable alternative for PRE called *Notarius*, a hybrid framework that decomposes PRE into two specialized steps: Overlapped NER Feature Extraction (ONFE) and Precedent Feature Aggregation (PFA). The ONFE module identifies fine-grained entities that encode features of the same precedents. These extracted entities are then processed and properly aggregated by the PFA module, reconstructing a complete representation of each precedent reference, suitable for downstream legal applications. We experiment with an implementation of *Notarius* that integrates a state-of-the-art open-source Uni-

fied NER method, W²NER (Li et al., 2022), within the ONFE module, and domain-specific heuristics within the PFA module.

We evaluated *Notarius* using two strategies. First, in the development stage, we measured the ONFE module’s feature extraction quality and conducted a qualitative expert assessment to verify the accuracy and completeness of the final extracted references on a collection of judicial decisions. Second, we assessed robustness in a production environment, where experts continuously reviewed outputs from real-world documents. This approach allowed us to identify edge cases that may not appear in curated datasets and to benchmark *Notarius* against the previous LLM-based pipeline in terms of reference-level quality, cost, and span-level offset traceability.

Our contributions are summarized as follows: i) We formalize PRE as a feature-binding problem and introduce *Notarius*, a cost-effective, domain-aware approach to PRE that reconstructs full precedent references through sequential feature extraction and aggregation; ii) We present key insights derived from a full-scale industrial migration from a generative LLM pipeline to a local deployment of *Notarius* combining a state-of-the-art NER model with domain-specific heuristics, enabling span-level offset traceability and eliminating continuous prompting overhead; iii) We validate *Notarius* through two complementary evaluations: a post-training qualitative assessment conducted by legal experts on a curated dataset of Brazilian judicial decisions, and continuous expert-driven monitoring of the system in production using real user documents.

2 Related Work

2.1 Legal NER and Precedent Extraction

Initial research in Brazilian Portuguese Legal NER focused primarily on identifying broad legal and universal entities. The foundational LeNER-Br dataset (Luz de Araujo et al., 2018) maps specific legal entity types, including precedents, across judicial decisions. However, its annotation scheme is strictly restricted to flat entities. Similarly, the Ulysses project (Câmara dos Deputados, 2018) significantly expanded available domain resources by introducing the UlyssesNER-Br (Albuquerque et al., 2022) and UlyssesLegalNER-Br (Albuquerque et al., 2026) corpora, built upon several types of legal documents. However, they also restrict annotation to flat entities, effectively not handling the overlapped or fragmented nature of precedent citations. Furthermore, a more granular perspective was introduced by Correia et al. (2022), which breaks down legal and precedent citations from documents from the Brazilian Supreme Federal Court (STF) into constituent parts. While this corpus is one of the few legal datasets to incorporate overlapping structures, its scope remains narrow, limited to the STF domain, and it stops short of the feature aggregation needed to map fragmented precedent components back to unique, resolvable entities.

More broadly, existing Legal NER and precedent extraction work has focused on either identifying precedents as flat entities or extracting their fine-grained components and, to the best of our knowledge, the reconstruction of complete precedent references from such fragmented features has not been directly addressed. Consequently, these approaches solve a different extraction problem and are not directly comparable baselines for PRE as formulated in this work.

2.2 Unified NER Methods

The limitations of traditional flat extraction have driven the development of diverse unified NER methods. Generative methods, such as BARTNER (Yan et al., 2021), recast entity extraction as a sequence-to-sequence problem. Among discriminative approaches, T2-NER (Huang et al., 2023) utilizes a two-stage entity-relation extraction setup, whereas W²NER (Li et al., 2022) formulates entity recognition as a word-word relation classification task. While these supervised architectures show high structural accuracy, their reliance on large an-

notated datasets is a current bottleneck due to the cost and complexity of annotating overlapped and discontinuous entities.

LLM-based unified NER methods, such as PromptNER (Ashok and Lipton, 2023), leverage zero or few-shot prompting to extract complex entity lists dynamically. While these prompt-based generative methods are able to handle overlapping and discontinuous text without local training, their long-term deployment in commercial environments reveals significant trade-offs. As mentioned in section 1, generative LLMs introduce recurring per-token API costs and fail to achieve the span-level offset traceability required by downstream structural applications. This makes highly optimized, non-LLM architectures a superior alternative for stable, localized feature extraction.

In summary, current literature presents a challenging dilemma for deployment-ready PRE. On one hand, supervised unified NER methods are constrained by a lack of publicly available, annotated datasets mapping fragmented precedent references, as described in subsection 2.1. On the other hand, adopting generative LLM-based methods to circumvent this data dependency introduces structural problems in production, namely recurring token-based financial overhead and a fundamental inability to provide exact span-level offset traceability.

3 The *Notarius* Framework

In this section we introduce *Notarius*, a novel PRE method developed to address the gaps identified in current literature, namely the lack of publicly available annotated legal datasets for fragmented precedent references, and the structural and financial limitations of LLM-based methods.

3.1 Overview

Notarius operates as a sequential pipeline. As illustrated in Figure 2, the process begins with raw legal text, which is first processed by the Overlapped NER Feature Extraction (ONFE) module. ONFE extracts the precedent citation strings together with overlapped and neighboring entities that encode additional features of the same precedent. These extracted elements are then passed to the Precedent Feature Aggregation (PFA) module, which clusters and consolidates related entities into structured precedent references.

3.2 Overlapped NER Feature Extraction (ONFE)

Since traditional flat NER models typically assume that entity spans are disjoint (Dai et al., 2020; Yan et al., 2021; Li et al., 2022), their applicability to PRE is limited, because precedent references frequently contain overlapped and discontinuous entities, as discussed in section 1. Our ONFE module partially addresses this limitation by incorporating Overlapped NER methods, which can identify both flat and overlapped entities. Consequently, the module can extract fine-grained entities that encode specific precedent features, together with the citation string. In the example presented in Figure 2, the ONFE module extracts the citation string and additional features of two different references from the input text. At this stage, however, ONFE does not determine which features belong to which specific reference, an essential step in documents that contain multiple, potentially interleaved precedent references.

3.3 Precedent Feature Aggregation (PFA)

If a citation string already contains all the information required to uniquely reference a precedent, no additional processing would be necessary, in principle. In practice, however, real-world legal texts often present this information in fragmented (discontinuous) ways that make it difficult for the ONFE module to recover a complete reference, even when all required features are present in the document. This challenge is amplified in industrial settings, where we cannot assume perfect citation string extraction accuracy or consistent authoring conventions.

The PFA module addresses both issues by performing feature aggregation over the entities extracted by the ONFE module, acting as a logical aggregation layer that resolves spatial and semantic ambiguity. The aggregation process is guided by a feature-binding framework that analyzes contextual proximity and structural dependencies between extracted citation strings and their associated precedent features. Based on these signals, PFA maps each feature to the correct precedent and consolidates them into a single structured reference, as illustrated in Figure 2. The process by which these heuristics were derived, from an analysis of expert-annotated citation patterns to iterative refinement on production errors, is detailed in subsection 4.2.

Although this feature-binding process is related

to traditional Relation Extraction (RE), the two differ in their objectives. RE typically predicts predefined semantic relations between entities, whereas PFA determines which heterogeneous precedent features belong to the same reference and aggregates them into a complete representation. Thus, we formulate PFA as a feature-binding problem rather than a semantic relation classification task.

4 Experimental Setup

4.1 Corpus and Annotation

Our corpus consists of 548 public Brazilian judicial decisions, reflecting diverse legal writing styles. Preprocessing involved HTML cleaning, text normalization, and tokenization using spaCy’s Brazilian Portuguese large model, *pt_core_news_lg*. To manage document length, we applied dynamic segmentation (100–150 tokens with 10–15 token overlap), yielding 6,874 chunks. We reduced this to 2,180 chunks by removing those without annotated entities and partitioned the remainder into training, development, and test sets.

Legal experts followed a standardized annotation protocol, labeling nine entity types essential for uniquely identifying precedents: *CITATION*, *CLASS*, *NUMBER*, *STATE*, *RAPPORTEUR*, *DATE*, *VENUE*, *PUB_LOC* (Publishing Location), and *COURT*. Due to operational constraints, each expert annotated a distinct subset of documents; thus, inter-annotator agreement was not measured.

Preliminary analysis revealed significant underrepresentation of some entity types, as shown in Table 1. To mitigate the risk of reduced recall for minority labels due to this imbalance, we employed an LLM-based data augmentation strategy on the training split only. Synthetic but legally coherent samples were generated to increase the frequency of such labels while preserving domain consistency. Specific prompt engineering details cannot be disclosed due to internal policies. Development and test splits remained unchanged to ensure fair evaluation.

4.2 Notarius Configuration

We implemented the ONFE module in an industrial setting by incorporating W²NER (Li et al., 2022), a state-of-the-art Unified NER method, which is particularly suitable for PRE as it handles the extraction of flat and overlapped entities within a single architecture. We trained the model for 30 epochs, selecting the best-epoch model using the develop-

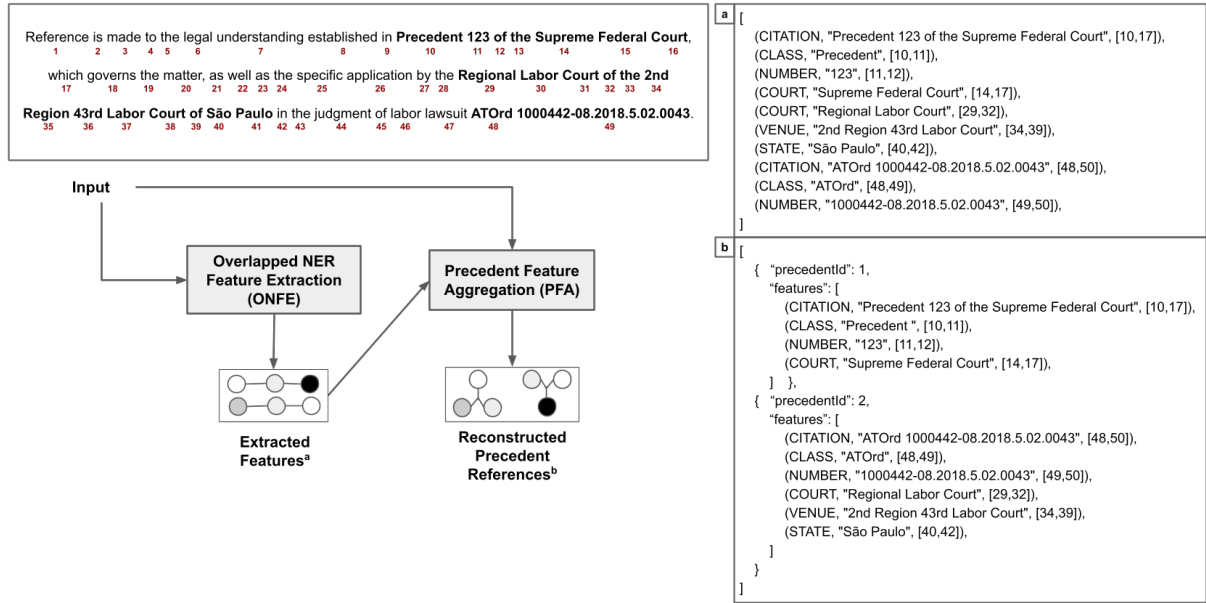


Figure 2: *Notarius* workflow.

ment split. To isolate the impact of data augmentation, we established two configurations under identical hyperparameters: a baseline model trained on the original distribution and an augmented model trained on the augmented training split described in [subsection 4.1](#).

For the PFA module, we designed a set of domain-specific heuristics that operate as a deterministic aggregation layer over the entities extracted by the ONFE module. These heuristics are designed to group the extracted features on the basis of their textual proximity to form candidate citations. To this end, we conducted an analysis of the expert-annotated data, which allowed us to segment citations into the following patterns: i) distant and isolated citations (all citations in [Figure 2](#)); ii) nearby and isolated citations (e.g., "...[Theme 465 of the STF] and [Precedent 123 of the TST]."); iii) agglutinated citations (e.g., "...did not incur direct and literal violation of [Themes 123, 456, and 789 of the STJ]"); and iv) nearby agglutinated citations (e.g., "[Themes 123, 456, and 789 of the STJ] and [Summaries 123 and 456 of the STF]"). These patterns also illustrate why textual proximity alone is insufficient for feature aggregation. Precedent features may occur non-contiguously or be interleaved with features associated with neighboring references, particularly in fragmented and agglutinated citation structures. Consequently, assigning each extracted feature to the nearest citation span may incorrectly bind features across adjacent precedents or split features belonging to the same prece-

dent. PFA therefore combines proximity with the structural dependencies captured by these citation patterns when reconstructing complete references. These patterns have guided the design of the heuristics that group the extracted labels according to their position in the text and expand the length of the obtained citation by looking at the other extracted labels in its neighborhood. We refined the heuristics iteratively, using errors observed in production to expose edge cases and citation styles absent from the curated corpus. Internal policies prevent further disclosure of the detailed rule logic and weights.

4.3 Evaluation Protocol

The evaluation framework mirrors the two-stage architecture of *Notarius*, assessing performance at two complementary levels. First, the **Feature-Level Evaluation** computes the F1-score of the ONFE module (baseline and augmented models) on the test set. Second, the **Reference-Level Evaluation** assesses whether *Notarius* extracts practically usable precedent references. Two legal experts (with a third for tie-breaking) manually inspected extracted references from 94 new documents outside the annotated corpus by answering: "Is the information in this reference sufficient to uniquely identify the cited precedent?" Responses were restricted to binary labels ("Yes" or "No"), and experts also recorded the ground-truth (actual) precedents. We then compute Precision (P), Recall

	Train	Aug. Train	Dev	Test	Feature-level (F1)		
					Base	Aug.	Δ
N. of Chunks	1533	1808	383	264	-	-	-
N. of Entities	9455	13024	2283	1437	-	-	-
Label	Distribution						
CITATION	21.47%	19.35%	22.38%	23.17%	0.57	0.56	-0.01
CLASS	17.44%	17.01%	17.13%	17.12%	0.71	0.68	-0.03
NUMBER	22.97%	20.91%	23.96%	23.87%	0.72	0.71	-0.01
STATE	3.97%	3.99%	3.59%	2.30%	0.81	0.86	+0.05
RAPPORTEUR	6.11%	8.12%	5.65%	4.52%	0.80	0.80	0.00
DATE	8.82%	10.07%	8.37%	5.64%	0.76	0.80	+0.04
VENUE	6.55%	7.71%	6.53%	8.35%	0.76	0.77	+0.01
PUB_LOC	2.74%	3.09%	2.93%	2.44%	0.83	0.80	-0.03
COURT	9.93%	9.74%	9.46%	12.60%	0.75	0.80	+0.05
Micro Avg F1	-	-	-	-	0.71	0.71	-
Macro Avg F1	-	-	-	-	0.75	0.75	-

Table 1: Dataset statistics and model feature-level extraction performance in F1-score (per entity type, and micro and macro averages).

(R), and F1-score as:

$$F_1 = \frac{2PR}{P+R}, \quad P = \frac{V}{E}, \quad R = \frac{V}{G},$$

where V , E and G denote the total counts of valid extracted references (“Yes” labels), references extracted by *Notarius*, and ground-truth references present in the documents, respectively.

5 Results

5.1 Feature-Level Evaluation

Table 1 reports the F1-scores for each feature entity type in the test set, for both the baseline and augmented models. A key observation is that both models achieved lower F1-scores for the *CITATION* label (0.57 and 0.56) than all other precedent features. This difference is expected because *CITATION* represents heterogeneous and frequently overlapping composite spans that may combine multiple legal elements into a single reference, whereas the remaining feature entities typically correspond to more localized and semantically homogeneous spans. This finding further supports the architectural decision to decompose PRE into sequential stages of feature extraction and aggregation. Furthermore, LLM-based data augmentation did not improve overall micro/macro average F1 (stagnant at 0.71 and 0.75). At the entity level, higher F1-scores were observed after augmentation for *STATE* (+0.05), *DATE* (+0.04), and *COURT* (+0.05). However, these per-class differences are

Total Ground-Truth (Actual) References (G)	428
Total Extracted References (E)	415
Total Valid References (V)	377
Precision (P)	0.91
Recall (R)	0.88
F1-score (F1)	0.89

Table 2: Expert inspection results for extracted references.

descriptive and should be interpreted cautiously given the limited number of evaluation instances for these entity types and that no statistical significance testing was performed.

5.2 Reference-Level Evaluation

Table 2 summarizes the manual inspection by legal experts. *Notarius* achieved a reference-level F1-score of 0.89, substantially higher than the F1-score obtained for the *CITATION* label alone. Although these scores correspond to different evaluation levels and do not constitute a controlled ablation of PFA, the result is consistent with the motivation for feature aggregation: citation spans alone may not contain all the information required to recover a complete precedent reference. By binding complementary extracted features, PFA can reconstruct references even when the citation string is missed or only partially extracted, provided that sufficient identifying information is available. Common invalid references include: (1) partial misses in com-

pound references (e.g., “SUMMARIES 362 and 382 from TST”), where the system binds features to only one precedent; and (2) insufficient metadata (e.g., “STF, on 13.11.2014”), which lacks other required features.

6 Deployment and Industrial Impact

In production at Jusbrasil, *Notarius* replaced a pure LLM-based pipeline. During the production evaluation, both systems were evaluated on the same set of production documents, over the same evaluation timeframe, and using the Reference-Level Evaluation protocol described in [subsection 4.3](#). Under this controlled comparison, *Notarius* achieved an F1-score of 0.92, compared with 0.89 for the legacy LLM pipeline. The legacy pipeline used few-shot prompting with ChatGPT-4o (snapshot gpt-4o-2024-08-06) to extract precedent references from the input text. The complete prompting strategy cannot be disclosed due to company intellectual property policies. Importantly, the LLM baseline F1-score of 0.89 reported here is independent of the identically valued 0.89 F1-score reported for *Notarius* in [subsection 5.2](#), which was obtained from a separate evaluation on the 94 documents outside the annotated corpus described in [subsection 4.3](#).

The two systems also differ fundamentally in their operational cost structure. The LLM baseline incurs a recurring API cost for every processed document, determined by the number of input and output tokens and the public API pricing of ChatGPT. Consequently, its API cost scales with the volume of processed documents. *Notarius*, in contrast, runs a smaller discriminative model on in-house infrastructure and incurs no per-document third-party API cost. Its operational cost instead consists of local inference and maintenance overhead, which is amortized over the production workload. At production scale, replacing recurring per-document API charges with local inference substantially reduces operational expenditure. Absolute production costs cannot be reported due to company confidentiality policies. Importantly, this cost difference is independent of extraction quality: the higher F1-score achieved by *Notarius* does not itself account for the reduction in operational costs. The two systems differ both in predictive performance and in deployment architecture. Moreover, although *Notarius* requires an upfront investment in data annotation and model development, these costs do

not recur for every processed document, whereas the LLM baseline continues to incur token-based API charges throughout inference.

Beyond efficiency, the discriminative architecture of *Notarius* enables direct span matching through entity offsets, unlike generative methods, allowing full traceability between each extraction and its source text, which supports downstream tasks such as automated text highlighting and dynamic hyperlinking. *Notarius* also produces consistent outputs for identical inputs, ensuring determinism and reproducibility, which are critical for legal auditability.

7 Conclusion

We presented *Notarius*, a hybrid framework for Precedent Reference Extraction designed to address the financial and structural limitations of generative models in continuous production environments by decomposing the task into sequential stages of feature extraction (ONFE) and aggregation (PFA). Our implementation using W²NER and domain-specific heuristics achieved an F1-score of 0.89 in a manual legal expert evaluation, extracting accurate references even when precedent features are dispersed and citation strings are fragmented. The industrial deployment of *Notarius* demonstrates that shifting from an LLM-based method to a local, discriminative alternative eliminates recurring third-party API costs while providing the span-level offset traceability required by downstream applications. More broadly, this work suggests that hybrid discriminative pipelines remain essential for high-precision structured extraction tasks even in the era of large generative models.

Limitations

While *Notarius* demonstrates high precision and significant operational efficiency in a production environment, some limitations should be considered when evaluating our pipeline. First, due to resource constraints in the industrial deployment timeline, each legal expert annotated a distinct, non-overlapping subset of the document corpus. Consequently, inter-annotator agreement was not measured, meaning that, while the experts followed a highly standardized and rigorous protocol, we could not statistically quantify the level of subjective bias or annotation drift across different annotators.

Additionally, the PFA module relies on a deterministic layer of heuristics tailored to internal corporate IP guidelines. Because these specific rule structures are proprietary to Jusbrasil and cannot be fully disclosed, the system is only partially reproducible by the research community, though the underlying sequential architecture remains generally applicable. A further limitation is the absence of a controlled comparison between PFA and simpler aggregation strategies. We did not retain gold feature-to-precedent assignments that would permit automatic evaluation of alternative aggregation methods. Moreover, our reference-level evaluation requires legal experts to assess whether each reconstructed reference uniquely identifies the cited precedent, and the expert annotation capacity required to evaluate the outputs of alternative aggregators is no longer available. Consequently, we cannot quantify how much PFA improves over simpler aggregation strategies, such as assigning each feature to the nearest citation span.

Furthermore, our evaluation and implementation are strictly bounded by the structural patterns and highly specific citation formatting of judicial decisions in the Brazilian civil-law system. Therefore, generalization of this framework to common-law jurisdictions, other languages, or other legal document types, such as legislative acts, government regulations, or contracts remains unexplored and would likely require further analysis at both the ONFE and PFA levels. We leave cross-jurisdictional, cross-language and cross-document-type evaluations to future work. Moreover, our evaluation design relies entirely on a binary judgment by legal experts regarding whether a reference contains sufficient metadata to uniquely identify a precedent, an approach that does not entirely capture granular nuances in extraction quality. Additionally, production monitoring was conducted exclusively within a single platform of Jusbrasil, meaning the robust performance reported may reflect the specific document distribution and user base of this environment rather than the entire universe of legal applications.

Finally, latency and throughput were not systematically measured during the production comparison and therefore cannot be reported retrospectively. Consequently, our operational comparison focuses on extraction quality and the structural difference in deployment costs rather than runtime performance.

Ethics Statement

Our work adheres to rigorous ethical standards regarding data privacy and the deployment of legal AI tools. The underlying corpus consists exclusively of public Brazilian judicial decisions that are accessible under national transparency laws. However, because Brazilian legal documents frequently contain raw Personally Identifiable Information (PII), the industrial production pipeline at Jusbrasil strictly implements internal data protection policies and sanitization routines to ensure compliance with data privacy regulations. To maintain complete evaluation integrity, our LLM-based data augmentation strategy was confined to the training split, while the development and test partitions remained entirely unaltered to prevent any synthetic bias from distorting our reported performance. Additionally, *Notarius* is engineered purely as an assistive legal research utility rather than an automated decision-making tool. Finally, the pipeline ensures auditability through *Notarius*'s span-level offset traceability, facilitating the verification and validation of all extracted source spans before they are used in downstream applications.

Acknowledgments

The authors thank the legal domain experts at Jusbrasil for their valuable contributions. This work was partially supported by Jusbrasil under the UFAM/Jusbrasil Doctoral Fellowship granted to Arthur Torres; by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Finance Code 001; by the Amazonas State Research Support Foundation (FAPEAM) through the POSGRAD 2025 program and by the NeuralBond Project (UNIVERSAL 2023, Proc. 01.02.016301.04300/2023-04); and by CNPq under INCT IAIA (406417/2022-9), INCT TILDIAR (408490/2024-1), and individual research grants to Altigran da Silva (301925/2025-9) and Edleno de Moura (310573/2023-8).

References

- Feisal Alaswad, E. Poovammal, and Batoul Aljaddouh. 2026. [From COCOMO to GPT: A Comprehensive Evaluation of LLM-Based Software Effort Estimation](#). *IEEE Access*, 14:36051–36075.
- Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P.

- Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. [UlyssesNER-Br: A corpus of Brazilian Legislative Documents for Named Entity Recognition](#). In *15th Computational Processing of the Portuguese Language (PROPOR 2022)*, pages 3–14.
- Hidemberg O. Albuquerque, Ellen Souza, Danilo C. G. Lucena, Héldon J. O. Albuquerque, Nádia F. F. da Silva, Márcio de S. Dias, Rafael O. Nunes, Adriano L. I. Oliveira, and André C. P. L. F. de Carvalho. 2026. [UlyssesLegalNER-Br: from Legislative to Legal, a comprehensive corpus of Brazilian legal documents for Named Entity Recognition](#). In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 331–341, Salvador, Brazil. Association for Computational Linguistics.
- Dhananjay Ashok and Zachary C. Lipton. 2023. [PromptNER: Prompting For Named Entity Recognition](#). *Preprint*, arXiv:2305.15444.
- Fernando A. Correia, Alexandre A.A. Almeida, José Luiz Nunes, Kaline G. Santos, Ivar A. Hartmann, Felipe A. Silva, and Hélio Lopes. 2022. [Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court](#). *Information Processing & Management*, 59(1):102794.
- Câmara dos Deputados. 2018. Ulysses: Robô de Inteligência Artificial. <https://www.camara.leg.br/noticias/548730-camara-lanca-ulysses-robo-digital-que-articula-dados-legislativos/>. Accessed: May-21-2026.
- Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. [An Effective Transition-based Model for Discontinuous NER](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.
- Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. 2023. [T2-NER: A Two-Stage Span-Based Framework for Unified Named Entity Recognition with Templates](#). *Transactions of the Association for Computational Linguistics*, 11:1265–1282.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. [Unified Named Entity Recognition as Word-Word Relation Classification](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:10965–10973.
- Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. [LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text](#). In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil. Springer.
- Hugo Mentzingen, Nuno Antonio, and Fernando Bação. 2025. [Effectiveness in retrieving legal precedents: exploring text summarization and cutting-edge language models toward a cost-efficient approach](#). *Artificial Intelligence and Law*, pages 1–21.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. [GPT-NER: Named Entity Recognition via Large Language Models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. [A Unified Generative Framework for Various NER Subtasks](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.
- Qi Zhan, Yile Wang, and Hui Huang. 2026. [Assessment of Generative Named Entity Recognition in the Era of Large Language Models](#). *Preprint*, arXiv:2601.17898.
- Johannes Zschache and Tilman Hartwig. 2026. [Comparing energy consumption and accuracy in text classification inference](#). *Scientific Reports*, 16:12717.