

CITAR-NER-Br: A Unified NER Dataset for Fine-Grained Legal Citation Extraction in Brazilian Portuguese

Arthur Elwing Torres^{†1}, Altigran Soares da Silva¹, Hidemberg O. Albuquerque²,
Ellen Souza Pereira², Héldon J. O. Albuquerque², Nádia Félix Felipe da Silva³,
Márcio de Souza Dias³, Douglas Vitorio⁴, André Spritzer⁵,
Laura P. Gouveia⁶, Rafael P. Gouveia⁶, Breno Gabriel de M. Lima⁷

¹Universidade Federal do Amazonas, ²Universidade Federal Rural de Pernambuco,

³Universidade Federal de Catalão, ⁴Instituto Federal de Pernambuco,

⁵Universidade Federal do Rio Grande do Sul, ⁶Universidade de São Paulo,

⁷Universidade Federal de Pernambuco

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

{hidemberg.albuquerque, ellen.amos, heldon.jose}@ufrpe.br

{nadia.silva, marciodias}@ufcat.edu.br

douglas.vitorio@pesqueira.ifpe.edu.br

{laura.gouveia, rafael.p.gouveia}@usp.br

bgml@cin.ufpe.br

[†]Corresponding author.

Abstract

We present CITAR-NER-Br, a unified NER dataset for direct extraction of fine-grained legal citation subcomponents in Brazilian Portuguese. CITAR-NER-Br annotates six citation subcomponents over 270 legal documents from LeNER-Br and Ulysses Tesemão, capturing the overlapping and discontinuous spans intrinsic to legal citations. To the best of our knowledge, this is the first unified NER dataset for the legal domain specifically designed to represent discontinuous entities, which account for 44.8% of annotations, the highest proportion among existing legal NER corpora. We establish baselines using four configurations across three LLM-based and non-LLM NER methods to support future research. Our experiments demonstrate that, while non-LLM methods achieve strong performance, they remain heavily reliant on high-quality training data. Conversely, LLM-based methods achieve overall lower performance, but require only few-shot supervision. Moreover, the traditional exact-match NER evaluation approach significantly penalizes LLM-based methods even when their outputs preserve semantic accuracy.

1 Introduction

The digitalization of the Brazilian judicial system has led to an unprecedented volume of legal documentation, requiring robust computational tools to parse and analyze complex legal texts. A fundamental challenge in this domain is the accurate extraction of legal citations (Câmara dos Deputados, 2018; Finatto and Macohin, 2026). In particular,

the Named Entity Recognition (NER) task, used for identifying and classifying textual references into categories (Santos et al., 2006; Li et al., 2022), has been increasingly used in the legal field (Souza et al., 2026). This task enables downstream applications such as automated legal search (Belandi and Siccardi, 2025), legal document segmentation (da Silva et al., 2022), and the construction of high-fidelity knowledge graphs (Ngandu et al., 2025). Additionally, by decomposing these citations into their constituent parts, legal information systems can automatically cross-reference legal elements across vast, heterogeneous legal documents, enabling retrieval and analysis at the level of individual citation subcomponents.

Current research in Brazilian legal NER largely focuses on identifying broad legal components, such as individual Laws and Precedents, and universal entities, such as Persons and Organizations, most of them based on HAREM (Santos et al., 2006), one of the first works in Portuguese NER. For the extraction of legal citations, the literature reveals two predominant approaches which we term: **Single-entity**, which extracts the entire legal citation as a single, joint entity, or **Citation-fragments**, which decomposes the citation into fragmented, disjointed entities. The example in Figure 1 illustrates this distinction using a citation containing multiple subcomponents. Under the Single-entity approach (Figure 1a), these components are captured together as one entity (“Article 51, paragraph 2, of the Organic Law of the Federal District”), whereas the Citation-fragments

approach (Figure 1b) identifies the *ARTICLE* (“Article 51”), *PARAGRAPH* (“paragraph 2”), and *LAW* (“Organic Law of the Federal District”) as isolated spans. Both approaches require post-processing: decomposition of monolithic entities for the former, and merging of ambiguous fragments for the latter, thereby falling short of supporting the direct extraction of resolvable citation subcomponents.

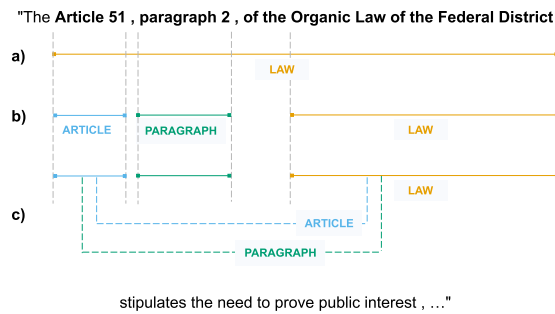


Figure 1: Translated snippet of CITAR-NER-Br showing multiple approaches for legal citation NER: (a) Single-entity, (b) Citation-fragments, and (c) Direct citation subcomponents. The example highlights how increasingly specific legal subcomponents incorporate the contextual information required for their resolution.

Considering these limitations, we present **Direct citation subcomponents**, a novel approach that extracts each citation subcomponent independently while preserving the full context to ensure accurate resolution. As illustrated in Figure 1c, the *LAW* entity (“Organic Law of the Federal District”) is extracted directly, while the *ARTICLE* entity (“Article 51 Organic Law of the Federal District”, a discontinuous span in which the intervening text is omitted) incorporates the *LAW* context required to identify the legal provision to which it refers. Furthermore, the *PARAGRAPH* entity (“Article 51 paragraph 2 Organic Law of the Federal District”) incorporates both the *ARTICLE* and *LAW* contexts in order to be fully resolvable. This representation naturally produces overlapping and non-contiguous spans, motivating the tasks of Overlapped and Discontinuous NER (Li et al., 2022; Yan et al., 2021). Although this example provides a simple illustration of these structures, Brazilian legal documents can contain substantially more complex and interleaved citations. Figure 2c illustrates such a case, where all citation subcomponents overlap with each other, and every entity other than “Federal Constitution” (*LAW*) is also discontinuous.

The task of jointly extracting flat, overlapped, and discontinuous entities into a single, cohesive framework is known as Unified NER (Li et al.,

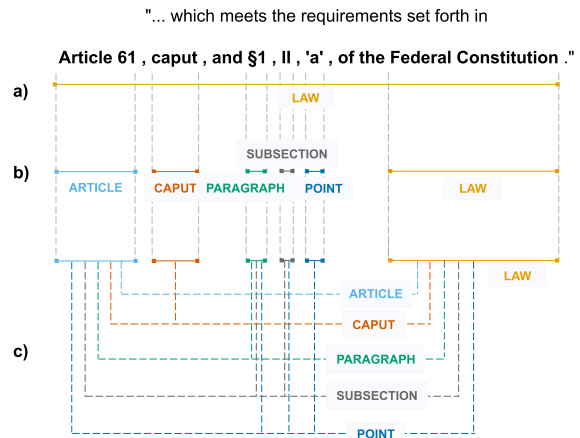


Figure 2: Translated snippet of CITAR-NER-Br illustrating the same three approaches to legal citation NER in a more complex setting: (a) Single-entity, (b) Citation-fragments, and (c) Direct citation subcomponents. The example highlights the highly interleaved, overlapping, and discontinuous structure of complex legal citations.

2022; Yan et al., 2021). In contrast to standard NER, which typically assumes that entities correspond to single contiguous spans with non-overlapping boundaries, Unified NER relaxes these structural constraints, allowing entities to overlap and to comprise multiple non-contiguous spans. Accordingly, our proposed approach is not standard sequence-labeling NER: a single textual segment may participate in multiple entities, and the spans composing an entity need not be adjacent.

Although unified NER can capture the inherent complexities of human language, it has received limited attention in the broader NLP literature. Moreover, the effectiveness of unified NER methods is fundamentally tied to the quality and structure of the underlying data, particularly when dealing with the intricate patterns of overlap and discontinuity. Even in the era of Large Language Models (LLMs) with zero or few-shot techniques (which do not require high-volume data for deployment), a high-quality, domain-specific benchmark dataset is necessary for rigorously evaluating and comparing the performance of NER methods.

To address the data availability gap, we present CITAR-NER-Br (Citation Information Text Annotation Resource), a novel Brazilian legal unified NER dataset specifically designed for the direct extraction of legal citation subcomponents. CITAR-NER-Br comprises 7,540 entities annotated in 270 legal documents, sourced and expanded from two previously published legal corpora, LeNER-Br (Luz de Araujo et al., 2018) and Ulysses Tesemõ (Siqueira et al., 2024). To the

best of our knowledge, this is the first legal unified NER dataset, and the second unified NER dataset in Brazilian Portuguese. In addition, 44.8% of the entities in our dataset are discontinuous, the highest reported proportion for such entities. CITAR-NER-Br was annotated by nine computer science researchers and was further refined through a post-hoc LLM-based verification process to ensure high-fidelity labeling. Finally, using CITAR-NER-Br, we implement and evaluate four configurations across three state-of-the-art NER methods, including both LLM-based and non-LLM approaches, thereby establishing baselines to support future research in this field. Our experiments show that the non-LLM supervised methods achieve the strongest performance, although they are heavily dependent on high-quality training data, which is costly to obtain given the high cognitive load of annotating complex discontinuous entities. Conversely, both configurations of the LLM-based method yield worse overall performance but offer a training-free, few-shot alternative. Furthermore, our analysis reveals that traditional exact-match NER evaluation heavily penalizes LLMs for minor textual variations even when their outputs preserve semantic accuracy, suggesting a need for more flexible evaluation criteria.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the annotation process. Section 4 presents the dataset and statistics. Section 5 describes the experiments and discusses the results. Section 6 concludes the paper and outlines future directions.

2 Related Work

2.1 Legal NER in Brazilian Portuguese

Research in Brazilian Portuguese legal NER has traditionally focused on identifying broad legal and universal entities in legal documents. The LeNER-Br dataset (Luz de Araujo et al., 2018), a foundational benchmark in this area, provides annotations with universal and specific legal entities in 70 documents, among which are 66 legal decisions from several Brazilian Courts alongside four specific legislation documents. The annotation scheme used is restricted to flat entities, with law annotations following the Single-entity approach.

The Ulysses project (Câmara dos Deputados, 2018) presents the UlyssesNER-Br corpus (Albuquerque et al., 2022) with 950 legislative documents, comprising public bills (PL-corpus), in-

ternal documents (ST-corpus) and public comments (C-corpus) (Costa et al., 2022), applying universal and domain entity types. This corpus was later expanded to UlyssesLegalNER-Br (Albuquerque et al., 2026) by leveraging Ulysses Tesemão (Siqueira et al., 2024), an unannotated collection of 3.5 million legal texts. The extended corpus adds new entities and introduces new corpora, for case law (Juris-corpus), and the first NER corpus exclusively with laws (Leis-corpus). All UlyssesNER corpora follow the Single-entity approach and are restricted to flat entities.

Correia et al. (2022) present a more granular perspective using documents from the Brazilian Supreme Federal Court (STF). Their corpus focuses on law and precedent citations, academic references and person names, extracting the main citations along with their intrinsic subcomponents. Their annotation scheme follows the Citation-fragments approach, breaking law citations into constituent parts. This work represents a significant milestone as it is one of the few legal datasets to incorporate both flat and overlapping entities, acknowledging the hierarchical legal nature. However, their approach remains limited to the STF domain and also stops short of directly extracting resolvable citation subcomponents.

Beyond dataset construction, recent works have proposed improvements to legal NER applications. Garcia et al. (2024) present the RoBERTaLexPT model, demonstrating that domain-adaptive pre-training on mixed data outperforms larger generic models in legal NLP tasks, including NER. Nunes et al. (2024) improved NER performance on the PL-corpus using a transformer-based self-learning approach. Previous work has also highlighted the applicability of NER for downstream legal tasks, such as Anselmo and Ribas (2025), which pseudonymize sensitive data in legal documents.

2.2 Unified NER datasets

Unified NER datasets, which encompass flat, overlapping, and discontinuous entities, have seen the most significant development within the biomedical domain. Representative datasets include Share13 (Pradhan et al., 2013), Share14 (Mowery et al., 2014), and CADEC (Karimi et al., 2015), providing annotations of drug reactions and clinical symptoms. In these benchmarks, discontinuous entities typically account for approximately 10% of the total annotations.

BioNestedNER (Schneider et al., 2026) repre-

sents, to our knowledge, the first unified biomedical NER dataset in Brazilian Portuguese, addressing the complexity of nested and discontinuous clinical entities. However, its proportion of discontinuous entities is significantly lower than the average in the literature ($\approx 3\%$), demonstrating the need for new corpora that expand the limits of structural complexity. We use these works as a foundation for the legal domain application, introducing the first legal unified NER dataset with a significantly higher proportion of discontinuous entities.

2.3 Unified NER Methods

Unified NER has driven diverse methods. Generative approaches, such as BART-NER (Yan et al., 2021), treat NER as a sequence-to-sequence generation task. W²NER (Li et al., 2022) formulates the task as a word-word relation classification task, while T2-NER (Huang et al., 2023) formulates it as a two-stage entity-relation extraction. However, these supervised methods rely heavily on large annotated datasets. To mitigate this data reliance, recent work explores LLMs for unified NER. GPT-NER (Wang et al., 2025) relies on in-line tagging methods, involving the entities with special characters. PromptNER (Ashok and Lipton, 2023) utilizes a comprehensive prompt design to extract entities as a list. Unlike in-line tagging, this method natively handles overlapping and discontinuous entities in zero and few-shot settings. In this work, we select BART-NER, W²NER, and PromptNER as baselines to implement CITAR-NER-Br.

3 Annotation Setup

3.1 Data Collection

To build CITAR-NER-Br, we used the entire LeNER-Br corpus, selected as our foundational source as it is the most cited Brazilian legal NER dataset in the literature. To enhance the diversity of judicial styles, we further enriched our base using 200 documents from the Tesemõ corpus, based on its diversity of open legal documents. The documents were selected from two distinct Brazilian courts, *STJ* and *TRF5*, with 100 documents from each, ensuring a balanced representation of the legal linguistic variation. These documents were strategically selected because they feature shorter court decisions than those typically found in LeNER-Br, while still containing legal citations as required for this study (see subsection 4.1). In the following sections, we refer to these specific

subsets as UT-STJ and UT-TRF5. Because the available files from both LeNER-Br and Ulysses Tesemõ were already preprocessed, no further text cleaning or preprocessing steps were required.

3.2 Entity Types

The entity scheme for CITAR-NER-Br was derived from an extensive empirical analysis of legal citations present in the source documents. We identified six granular law subcomponents that are essential for resolving and cross-referencing legal mandates. Figure 2c provides visual examples of each entity type in context. A brief description of these entities is presented in Table 1. These entities typically follow a Brazilian legal hierarchy (Brasil, 1998), presented below:

$$\text{Law} \rightarrow \text{Article} \rightarrow \left\{ \begin{array}{l} \text{Caput} \\ \text{Paragraph} \end{array} \right\} \rightarrow \text{Subsection} \rightarrow \text{Point}$$

Entity Type	Description
LAW (<i>LEGISLAÇÃO</i>)	Comprises only Brazilian legal acts with power of law, including Law (<i>Lei</i>), Legislation (<i>Legislação</i>), Constitution (<i>Constituição</i>), Decree-Law (<i>Decreto-lei</i>) and Constitutional Amendment (<i>Emenda Constitucional</i>).
ARTICLE (<i>ARTIGO</i>)	The primary organizational unit of a law.
CAPUT	The opening text of a legal article.
PARAGRAPH (<i>PARÁGRAFO</i>)	A secondary division providing specific details, exceptions, or clarifications to an article.
SUBSECTION (<i>INCISO</i>)	Usually represented by Roman numerals, these elements break down the contents of an article or paragraph. For structural consistency in our schema, this entity type also encompasses ITEM.
POINT (<i>ALÍNEA</i>)	Lowercase lettered divisions typically nested within a subsection.

Table 1: Taxonomy and description of the six granular law subcomponents annotated in CITAR-NER-Br.

3.3 Annotation Guidelines

Given the high cost of recruiting legal experts for large-scale annotation, we relied on computer science researchers from our team. This choice was motivated by the highly structured nature of the annotation task, as legal citations feature highly structured patterns (e.g., Roman numerals for subsections; explicit anchors like “Article”) that allow non-experts to perform the annotation reliably following appropriate guidelines and training, as reflected in the agreement scores reported below. The team comprised nine native Brazilian Portuguese speakers from public federal universities (four of

whom had previous legal annotation experience), and included two doctoral-level and one master’s-level professors, one researcher with a PhD degree, three doctoral students, and two undergraduate students.

Following a preliminary training phase to align guidelines, annotations were executed over three cycles, with the first covering 20% of the data set to get early feedback on the annotation workflow, and the next two covering 40% each. Each cycle consisted of three phases: annotation, Inter-Annotator Agreement (IAA) measurement, and curation. To measure IAA, we follow previous work and adopt F_1 -score (Brandesen et al., 2020). Chance-corrected agreement metrics such as Cohen’s Kappa require defining the negative class, which is problematic for span-based NER because the set of possible non-entity spans is extremely large. Additionally, Hripcsak and Rothschild (2005) show that, when true negatives are abundant, the F-measure approaches chance-corrected agreement. Furthermore, we compute F_1 under strict exact-match criteria, such that any difference in entity type or span boundaries is treated as a disagreement. The scores in Table 2 show consistent agreement, ranging from 0.73 to 0.89, with minor drops in the UT-STJ collection due to a lower volume of entities, making its scores more sensitive to individual disagreements.

Collection	Phase 1			Phase 2			Phase 3		
	Pairs			Pairs			Pairs		
	1	2	3	1	2	3	1	2	3
LeNER-Br	0.80	0.79	0.81	0.86	0.77	0.84	0.69	0.83	0.85
UT-STJ	0.79	0.65	0.92	0.92	0.61	0.68	0.82	0.72	0.87
UT-TRF5	0.84	0.70	0.76	0.97	0.80	0.98	0.85	0.85	0.85
Full	0.81	0.75	0.81	0.89	0.73	0.82	0.78	0.81	0.85

Table 2: Pairwise IAA (F_1) per cycle, for the full dataset and constituent collections.

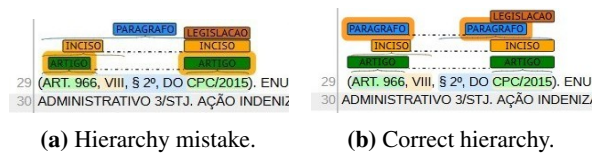


Figure 3: Snippet of CITAR-NER-Br annotation in BRAT.

Following the IAA measurement phase, each curator reviewed the annotations of its designated pair to resolve disagreements and correct overlooked errors. Finally, before advancing to the subsequent cycle, the entire team convened to analyze Inter-Annotator Agreement (IAA) trends and establish refined heuristics against recurring errors. All annotations were performed using BRAT (Stenetorp

et al., 2012)¹, as it remains, to our knowledge, the only open-source tool that supports discontinuous entities.

3.4 Hierarchy Errors and LLM Verification

In the Brazilian legal system, the standard hierarchy in subsection 3.2 is occasionally violated in some legislative texts. For example, one annotator initially made an incorrect annotation (Figure 3a), assuming that “§2º” (PARAGRAPH) was nested within “VII” (SUBSECTION). However, verification of the “CPC/2015” (the Brazilian Code of Civil Procedure) reveals that these components actually occupy the same hierarchical level. The corrected annotation (Figure 3b) reflects this specific law’s unique internal logic. Accurately identifying such cases would require extensive prior knowledge or manual verification of official sources for each citation. Given the corpus scale, requiring these external lookups for every entity would render the annotation process impractical.

As an additional verification, we employed GPT-4o (gpt-4o-2024-08-06) to review all annotations for correctness and resolvability. Chosen for its balance between reasoning ability and cost-effectiveness, this LLM model was tasked with identifying span errors and checking whether each entity provided sufficient context for resolving a specific law or subcomponent. Specifically, we prompted the LLM with the following question: *In Brazilian legislation, does [ENTITY] exist? Is this legal item a [ENTITY_TYPE]?²*. We also added few-shot examples (see Figure 6 in Appendix A for more details). Temperature was set to 0 to mitigate model randomness. To evaluate the efficacy of this cooperative step, a senior annotator manually reviewed every LLM response.

Collection	FP	TP (Changed)	TP (Unchanged)
LeNER-Br	948	86	262
UT-STJ	56	0	39
UT-TRF5	44	1	45
Full	1,048	87	346

Table 3: LLM verification step performance in False Positives (FP), True Positives with corrections (TP Changed), and True Positives left intact (TP Unchanged) across the full dataset and constituent collections.

Table 3 details the LLM-based verification performance, revealing that the model generated a high

¹<https://brat.nlplab.org/>

²Translated here for clarity; original prompt in Portuguese.

volume of False Positives (FPs), accounting for 70.8% of responses. However, the strategy successfully uncovered 87 concrete corrections (5.9% TP Changed) and 346 valid issues where annotations could not be modified (23.3% TP Unchanged). The main constraints preventing these 346 corrections were temporal context (198 occurrences, 57.2%), where verified entities existed historically at the time of document drafting, followed by original typographic or spelling errors in the legal document (139 occurrences, 40.2%).

4 CITAR-NER-Br Dataset

This section presents the descriptive statistics and structural properties of CITAR-NER-Br. To quantify linguistic characteristics, we perform tokenization using the `pt_core_news_lg` model from the spaCy library³. All annotation files are provided in the standard BRAT span-character format. To ensure transparency and reproducibility, the dataset and the full base code are publicly available⁴.

4.1 Statistics

The corpus exhibits significant diversity in document length and entity density. As shown in Figure 4a, documents in the LeNER-Br collection are substantially longer than those in the UT-STJ and UT-TRF5 collections, with some documents exceeding 20,000 tokens. This pattern is mirrored in Figure 4b, where the number of entities per document peaks in the LeNER-Br collection, often reaching hundreds of annotations per file. Furthermore, Figure 4c reveals that, while most entity spans are relatively concise (under 10 tokens), there is a notable tail of longer, complex citations.

Figure 4d illustrates the presence of discontinuous entities per document, where LeNER-Br maintains the highest density. Additionally, Figure 4e shows the distance between fragments of discontinuous entities, measured in tokens, on a logarithmic frequency scale. While the median fragment distance is 9.0, the LeNER-Br collection contains extreme outliers where fragments of a single entity are separated by over 5,000 tokens, posing a significant challenge for NER methods that assume entities are formed by adjacent tokens. Finally, an analysis of surface-form diversity reveals unique-to-total ratios ranging from 62% to 80% for all entity types except *LAW*, which has only 705 unique

surface forms among 2,740 annotations (26%).

4.2 Data Split

We partitioned the annotated dataset into training, development, and test sets using a 70/10/20 split. Given the extreme variability in document lengths, a simple document-count split would have resulted in highly imbalanced sets. Instead, we performed the split based on the total number of tokens to ensure proportional representation. Furthermore, we employed a stratified approach within each collection across all splits to maintain a consistent distribution of entity types both for all entities and for discontinuous entities only (details in Table 4).

Figure 5 illustrates the split strategy, showing highly stable proportions of entity types across the train, dev, and test sets for every collection, including when considering only discontinuous entities. This consistency supports that test set performance reflects generalization across the full corpus. Furthermore, we verified that there was no document leakage, ensuring that no document appeared more than once across the dataset splits.

Collection	Split	Docs	Tokens	Unique Tokens	Entities	Discont. Entities
LeNER-Br	train	55	230,382	17,256	4,893	2,288
	dev	7	38,022	5,879	518	242
	test	8	62,419	7,606	919	333
UT-STJ	train	79	37,076	5,378	462	260
	dev	11	6,104	1,775	87	40
	test	10	9,971	2,703	86	49
UT-TRF5	train	69	26,714	5,726	399	107
	dev	12	4,161	1,548	61	15
	test	19	7,336	2,390	115	44
Full	train	203	294,172	22,031	5,754	2,655
	dev	30	48,287	7,585	666	297
	test	37	79,726	10,109	1,120	426

Table 4: Statistics across the splits of the full dataset and its constituent collections.

4.3 Document Segmentation

The extensive length of legal documents presents a computational bottleneck for most transformer-based NER methods due to their strict input length limits. Due to this constraint, all documents were segmented into chunks of up to 200 subtokens using the BERTimbau tokenizer⁵ (Souza et al., 2020). To ensure consistency, the BERTimbau tokenizer was used solely for segmentation, while the spaCy tokenizer was reserved for descriptive statistics.

³<https://spacy.io/>

⁴<https://github.com/arttorres0/citar-ner-br>

⁵<https://huggingface.co/neuralmind/bert-base-portuguese-cased>

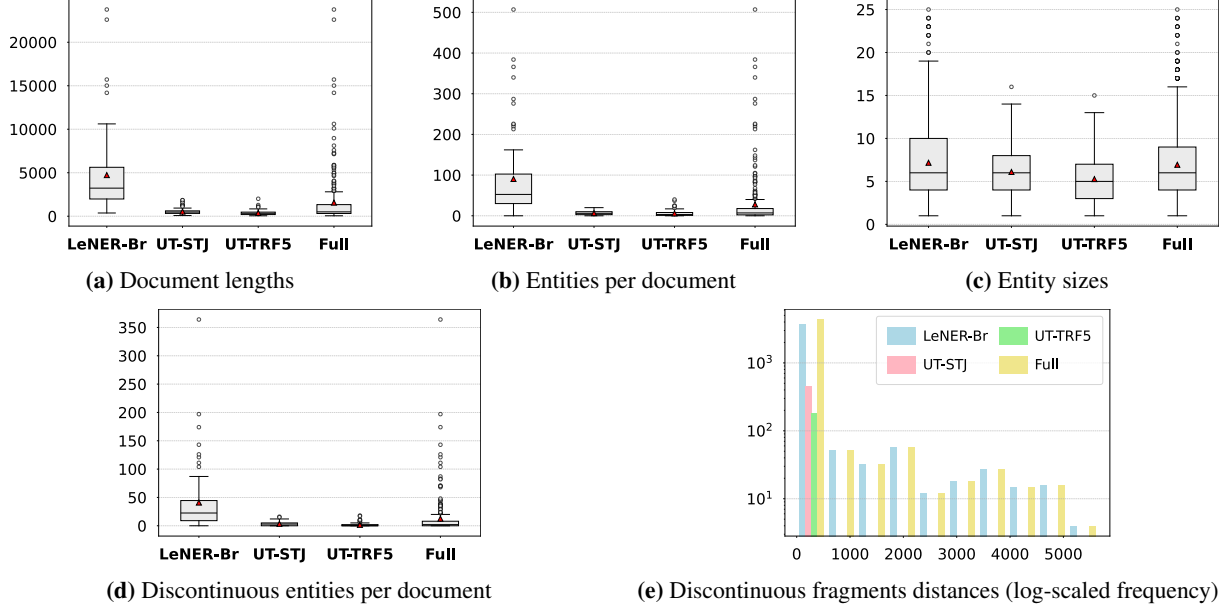


Figure 4: CITAR-NER-Br statistics. Lengths, sizes and distances are measured in tokens.

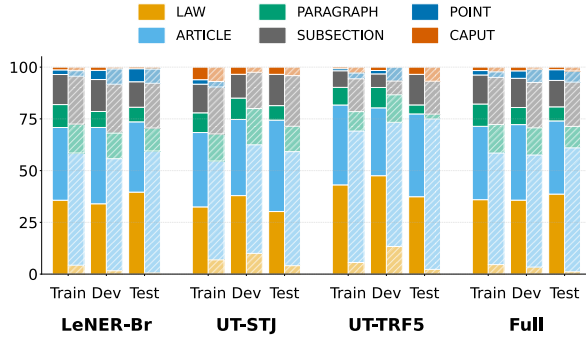


Figure 5: Entity distribution by dataset split and constituent collection. Saturated (left) bars denote total entities; desaturated (right) bars denote discontinuous entities only.

This constraint is particularly problematic for long-distance discontinuous entities: when their fragments fall into different chunks, the complete entity cannot be represented within a single model input and its annotation must therefore be discarded. As shown in Figure 4e, although the median distance between discontinuous fragments is 9.0 tokens, extreme cases span over 5,000 tokens.

Collection	Train	Dev	Test
LeNER-Br	1,623 (429)	258 (0)	437 (76)
UT-STJ	305 (6)	47 (0)	75 (4)
UT-TRF5	243 (1)	41 (0)	66 (13)
Full	2,171 (436)	346 (0)	578 (93)

Table 5: Total chunks across the splits of the full dataset and its constituent collections. Values in parentheses indicate the number of discarded annotations.

Table 5 summarizes the resulting chunk distribution and the total of discarded annotations. Trunca-

tion affects LeNER-Br (8.7% in the training split and 8.3% in the test split), due to its longer fragment distances, whereas UT-STJ and UT-TRF5 are minimally impacted. This segmentation and the resulting annotation removal apply exclusively to non-LLM baselines, whereas the LLM-based baselines used the full documents and all entities.

5 Experiments

This section presents the evaluation of BART-NER, W²NER, and PromptNER, three state-of-the-art unified NER methods, implemented on CITAR-NER-Br. Performance is measured using the traditional exact-match approach, requiring a perfect alignment between predicted strings and gold entity spans. The results comprise precision, recall, and F_1 -score, and are generated via a custom, publicly available evaluation script found in the CITAR-NER-Br repository. Furthermore, while dataset statistics are detailed per constituent collection to show structural differences, we report experimental results exclusively on the full dataset. This aligns with our training setup, where models are exposed to all collections simultaneously to robustly evaluate their generalization capabilities across the diverse spectrum of Brazilian judicial writing styles.

For BART-NER and W²NER, we used the original hyperparameters from their respective experiments on the CADEC, ShARe13, and ShARe14 datasets, with batch size reduced from 16 to 4 (GPU constraints) and W²NER epochs increased from 10 to 30. We used BERTimbau as the W²NER pre-

trained model. Training took 2h for W²NER and 11h40 for BART-NER on a single NVIDIA T4 GPU (16 GB VRAM, 16 vCPUs, 60 GB RAM).

For PromptNER, we follow the original prompt design but replace generic templates with domain-specific guidelines. We also include five stratified few-shot samples from the training split, augmented with negative entity samples per PromptNER guidelines. The template is in English (except for input texts and entity types) and is detailed in Figure 7 (Appendix A). We retain English for the prompt instructions to remain consistent with the original PromptNER setup and avoid introducing prompt language as an additional experimental variable. Nevertheless, the input documents and extracted entities remain in Brazilian Portuguese, so the underlying NER task is performed entirely over Portuguese text. We evaluate using Gemini v3.1 Pro (gemini-3.1-pro-preview, temperature=0) and ChatGPT v5 (gpt-5-2025-08-07, temperature=1, the minimum value allowed).

Finally, evaluating PromptNER on discontinuous entities is challenging because its text-based output lacks explicit layout offsets. We classify a predicted entity as continuous if it appears as an exact substring of the input, and discontinuous otherwise; misclassifications of conceptually continuous entities with minor text variations were exceptionally rare in our experiments.

5.1 Quantitative Results and Analysis

Table 6 reports per-type and micro-average metrics. Evaluating all entities, BART-NER leads with a micro- F_1 of 0.86, followed by W²NER (0.73), while PromptNER substantially underperforms (Gemini 0.47, ChatGPT 0.31). On discontinuous entities, the gap widens: BART-NER stays consistent, with a micro- F_1 of 0.73, while W²NER drops to 0.55 due to weakness outside *ARTICLE*, and ChatGPT and Gemini yield near-zero scores.

These results reveal key insights into how distinct architectures process complex, multi-layered citations. W²NER’s high precision on *LAW* spans suggests that its grid-based word-word relation matrix effectively handles local boundary constraints for continuous legal entities, yet its drop in performance on discontinuous subcomponents indicates that word-word relation classification struggles with long-distance structural dependencies. Conversely, BART-NER’s autoregressive sequence-to-sequence design proves highly adaptable to discontinuous syntax, as formulating extraction as a

target text generation problem naturally bypasses physical token-distance constraints. Finally, the high contrast between the high precision and low recall of the LLM-based configurations highlights a conservative generation bias, showing that, while ChatGPT and Gemini can accurately extract highly visible subcomponents, they suffer from a severe tendency to omit dense legal entities entirely.

5.2 Error Analysis and Discussion

A closer inspection of the outputs of each unified NER method reveals various error profiles. For BART-NER, failures stem from occasional spelling variations and truncated or excess tokens during autoregressive text generation. For instance, it fails by predicting “Law 9.717/1998 art. 102, I, f, of the constitutional text” instead of “art. 102, I, f, of the constitutional text”, inadvertently adding excess tokens. Despite these anomalies, its autoregressive architecture is the most effective in capturing the rigid syntax of Brazilian legal citations, achieving the most balanced performance. Conversely, W²NER’s grid classification schema relies heavily on local contextual consistency, leaving it prone to relation fragmentation. It struggles to build long-range classification chains for distant discontinuities. Additionally, occasional type misclassifications occur due to bias toward the initial tokens of a prediction string. For example, it misclassifies the prediction “art. 24, § 1º, of CF” as *ARTICLE*, when it is actually a *PARAGRAPH*. This explains its substantial performance drop in types other than *LAW* and *ARTICLE*. Overall, both BART-NER and W²NER performed well across the board, though at the cost of requiring large volumes of annotated data, a substantial challenge considering the extreme complexity of discontinuous entities.

Both PromptNER configurations highlight the steep limitations of unified NER LLM-based methods. Failures of PromptNER via ChatGPT stem from the extreme difficulty of identifying discontinuous entities. For instance, when analyzing “art. 102, I, f, of constitutional text”, it successfully extracts the continuous entities (*POINT* and *LAW*) but misses the discontinuous entities (*ARTICLE* and *SUBSECTION*). PromptNER via Gemini, on the other hand, demonstrates a much better capacity for generating discontinuous predictions. However, it is severely penalized by the traditional exact-match evaluation approach, which treats minor normalization shifts as complete misses, such as when predicting “art. 102 **of** constitutional text”,

Subset	Entity Type	BART-NER			W ² NER			PromptNER (ChatGPT v5)			PromptNER (Gemini v3.1)		
		P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
All Entities	LAW	0.93	0.94	0.93	0.96	0.96	0.96	0.90	0.28	0.43	0.89	0.52	0.65
	ARTICLE	0.87	0.86	0.87	0.59	0.82	0.69	0.40	0.16	0.23	0.40	0.29	0.33
	CAPUT	1.00	0.85	0.92	0.00	0.00	0.00	0.25	0.14	0.18	0.75	0.64	0.69
	PARAGRAPH	0.69	0.76	0.72	0.53	0.24	0.33	0.51	0.24	0.32	0.58	0.55	0.57
	SUBSECTION	0.72	0.70	0.71	0.42	0.20	0.27	0.36	0.17	0.23	0.32	0.25	0.28
	POINT	0.72	0.68	0.70	0.40	0.08	0.13	0.29	0.07	0.11	0.54	0.45	0.49
	Micro avg.	0.86	0.86	0.86	0.73	0.72	0.73	0.56	0.21	0.31	0.58	0.40	0.47
Discontinuous Entities	LAW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ARTICLE	0.85	0.80	0.82	0.62	0.75	0.68	0.04	0.02	0.02	0.11	0.09	0.10
	CAPUT	1.00	0.71	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.38	0.38
	PARAGRAPH	0.62	0.65	0.63	0.53	0.27	0.36	0.00	0.00	0.00	0.27	0.27	0.27
	SUBSECTION	0.62	0.62	0.62	0.36	0.25	0.29	0.00	0.00	0.00	0.04	0.03	0.03
	POINT	0.38	0.40	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.09	0.07
	Micro avg.	0.75	0.72	0.73	0.56	0.53	0.55	0.02	0.01	0.01	0.11	0.10	0.10

Table 6: Precision, Recall, and F_1 -score for each unified NER method across all entity types on the test set, evaluated for both all entities and discontinuous entities only.

with an excess token, instead of “art. 102 constitutional text”. This approach also penalizes the ChatGPT configuration when it introduces slight spelling variations, such as when predicting capitalized “Art. 16 of Law 8.443/1992” instead of “art. 16 of Law 8.443/1992”. Because these slight syntactic variations may preserve the underlying semantic accuracy, these systematic mismatches highlight a limitation of exact-match evaluation for generative NER. More flexible criteria could account for such surface-level variations through semantic-aware evaluation or normalization of predicted and reference entities. Another possibility is to link extracted entities to canonical legal references, allowing syntactically different predictions to be assessed based on whether they resolve to the same underlying entity. Furthermore, despite their overall worse performance than the non-LLM methods, they offer a key trade-off by requiring zero training and relying only on few-shot samples.

6 Conclusion and Future Work

In this work, we presented CITAR-NER-Br, a novel unified NER dataset explicitly designed for the direct extraction of fine-grained legal citation sub-components in Brazilian Portuguese. Unlike existing resources that rely on monolithic single-entity definitions or disjointed citation fragments, CITAR-NER-Br introduces a new paradigm by directly capturing overlapping and discontinuous legal elements. CITAR-NER-Br has the highest proportion of discontinuous entities (44.8%) among known unified NER benchmarks and fills a critical gap in legal NLP by providing a structured, domain-specific resource for evaluating state-of-the-art uni-

fied NER methods. To ensure high-fidelity labeling, we employed a cooperative LLM verification step to further refine the human annotations.

Our baseline experiments revealed clear performance trade-offs across distinct paradigms. Non-LLM methods, specifically BART-NER and W²NER, demonstrated strong capabilities in identifying the intricate boundaries of legal entities, though this robustness comes at the expense of requiring substantial volumes of high-quality training data, a costly process considering the complexities of annotating discontinuous entities. Conversely, the LLM-based prompting method PromptNER achieved lower overall performance scores while offering a training-free alternative that requires only a handful of few-shot samples.

These findings open two promising avenues for future research. First, given the underperformance of LLM-based methods, we aim to investigate specialized prompt architectures and in-context learning strategies tailored for unified NER, including Portuguese-language prompting and Portuguese-adapted LLMs, to improve the models’ ability to extract deeply interleaved and discontinuous entities. Second, our analysis underscores the limitations of the traditional exact-match evaluation approach in generative contexts. Because LLM-based methods typically introduce minor syntactic or spelling variations, this approach severely penalizes them. We intend to investigate more flexible evaluation criteria for generative unified NER, including semantic-aware and normalization-based approaches that account for structural alignment while remaining resilient to surface-level textual variations.

Limitations

The first limitation stems from the LLM-based verification process. Although it proved valuable for filtering and correcting annotations, it yielded a high number of false positives. Hence, future work will focus on refining this verification method by incorporating self-correction routines and more detailed prompting rules. This refinement will ensure greater precision in automated validation, ultimately laying the foundational basis for future annotation tasks to reliably deploy LLMs as cooperative annotation tools, especially in scenarios where domain-specific experts are unavailable or cost-prohibitive.

Another limitation concerns the generalization of our approach to other languages and legal systems. Legal citation conventions are closely tied to jurisdiction-specific structures and drafting practices, and the hierarchical patterns captured by CITAR-NER-Br reflect the characteristics of Brazilian legislation. Consequently, the proposed entity schema and its patterns of overlap and discontinuity may not transfer directly to jurisdictions with different citation conventions or legal structures. Cross-lingual transfer methods and domain-adaptive prompting provide promising directions for adapting the approach to these different settings.

Finally, despite BART-NER’s strong performance, extracting complex legal hierarchies directly remains a challenge. We plan to explore alternative architectures such as T2-NER, which reformulates unified NER as a two-stage entity-relation extraction task. Implementing this pipeline will require post-processing CITAR-NER-Br by transforming every discontinuous entity into a group of fragments bound by relational links. This programmatic conversion ensures that the dataset’s structural schema can be accurately mapped across the T2-NER pipeline while remaining completely traceable back to its original Direct Citation Sub-components design. By breaking down discontinuous citations into these independent fragments (Citation-fragments approach) and reconstructing them into discontinuous entities via a relational layer, this architecture can serve as a more robust and scalable parsing mechanism for legal citations in Brazilian Portuguese.

Ethical Considerations

The development and publication of CITAR-NER-Br adhere to strict ethical and legal standards, since

all source documents in CITAR-NER-Br are part of either LeNER-Br or Ulysses Tesemão, and both corpora are entirely publicly available and do not pose any data sensitivity concerns. Furthermore, the source documents were reused in compliance with their original licensing terms, as LeNER-Br is distributed under the MIT license and Ulysses Tesemão under the CC BY-NC 4.0 license. To support open science, CITAR-NER-Br is published under a CC BY 4.0 license. Regarding the annotation process, all members of the annotation team participated as volunteers, and no financial compensation was involved. Finally, the authors declare no conflicts of interest regarding the publication of this work.

Acknowledgments

This work was partially supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Finance Code 001; by Jusbrasil under the UFAM/Jusbrasil Doctoral Fellowship granted to Arthur Torres; 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 an individual research grant to Altigran da Silva (301925/2025-9).

References

- Hidemberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitória, 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.
- Marcelo Anselmo and Bruno César Ribas. 2025. *Pseudonymization in Legal Texts According to the*

- LGPD: A Named Entity Recognition Approach. In *Intelligent Systems*, pages 309–323, Cham. Springer Nature Switzerland.
- Dhananjay Ashok and Zachary C. Lipton. 2023. [PromptNER: Prompting For Named Entity Recognition](#). *Preprint*, arXiv:2305.15444.
- Valerio Bellandi and Stefano Siccardi. 2025. [Managing Legal Data: A Framework for Document Annotation and Retrieval](#). In *Proceedings of the 33rd Symposium on Advanced Database Systems*, volume 4182 of *CEUR Workshop Proceedings*, pages 604–614. CEUR-WS.org.
- Alex Brandsen, Suzan Verberne, Milco Wansleben, and Karsten Lambers. 2020. [Creating a Dataset for Named Entity Recognition in the Archaeology Domain](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.
- Brasil. 1998. [Lei Complementar nº 95, de 26 de fevereiro de 1998. Dispõe sobre a elaboração, a redação, a alteração e a consolidação das leis](#). Diário Oficial da União, Brasília, DF.
- 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.
- Rosimeire Costa, Hidemberg Oliveira Albuquerque, Gabriel Silvestre, Nádia Félix F. Silva, Ellen Souza, Douglas Vitorio, Augusto Nunes, Felipe Siqueira, João Pedro Tarrega, João Vitor Beinotti, Márcio de Souza Dias, Fabíola S. F. Pereira, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. [Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text](#). In *Progress in Artificial Intelligence*, pages 767–779, Cham. Springer International Publishing.
- 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.
- F. da Silva, G. Guimarães, R. Marcacini, A. Queiroz, V. R. P. Borges, T. Faleiros, and L. Garcia. 2022. [Named Entity Recognition Approaches Applied to Legal Document Segmentation](#). In *Anais do X Symposium on Knowledge Discovery, Mining and Learning*, pages 210–217, Porto Alegre, RS, Brasil. SBC.
- Maria José Bocorny Finatto and Aline Macohin. 2026. [PLN no Direito: Perspectivas e Desafios com Textos Jurídicos e Legais](#). In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 4 edition, volume 3, book chapter 12. BPLN.
- Eduardo Garcia, Nadia Silva, Felipe Siqueira, Juliana Gomes, Hidemberg O. Albuquerque, Ellen Souza, Eliomar Lima, and André de Carvalho. 2024. [RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 374–383, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- George Hripcsak and Adam S. Rothschild. 2005. [Agreement, the F-Measure, and Reliability in Information Retrieval](#). *Journal of the American Medical Informatics Association*, 12(3):296–298.
- 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.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. [Cadec: A corpus of adverse drug event annotations](#). *Journal of Biomedical Informatics*, 55:73–81.
- 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. de Oliveira, Matheus Staufer, 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.
- Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee M. Christensen, David Martínez, Liadh Kelly, Lorraine Goeuriot, Noémie Elhadad, Sameer Pradhan, Guergana K. Savova, and Wendy W. Chapman. 2014. [Task 2: ShARe/CLEF eHealth evaluation lab 2014](#). *Working Notes for CLEF Conference*, 1180:31–42.
- Bernard Ngandu, Jovita Mateus, Landry Mbale, and Antoine Bagula. 2025. [Automated Citation Detection in Congolese Legal Texts: Leveraging LLM-Based NER for Knowledge Graph Construction](#). In *2025 International Conference on Emerging Trends in Networks and Computer Communications (ET-NCC)*, pages 994–999.
- Rafael Oleques Nunes, Dennis Giovanni Balreira, André Suslik Spritzer, and Carla Maria Dal Sasso Freitas. 2024. [A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning](#). In *Proceedings of the 16th International Conference*

- on *Computational Processing of Portuguese - Vol. 1*, pages 290–300, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guer-gana K. Savova. 2013. [Task 1: ShARe/CLEF eHealth evaluation lab 2013](#). In *Working Notes for CLEF Conference*, volume 1179. CEUR-WS.
- Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. [HAREM: An Advanced NER Evaluation Contest for Portuguese](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Elisa Terumi Rubel Schneider, Yohan Bonescki Gumiel, Paloma Martínez, Claudia Moro, and Emerson Cabrera Paraiso. 2026. [BioNestedNER: A Hybrid Language Model Approach for Recognizing Nested, Discontinuous, and Multi-Type Named Entities](#). *Journal of the Brazilian Computer Society*, 32(1):635–648.
- Felipe Siqueira, Douglas Vitório, Ellen Polliana Souza, José Antônio Pedro Santos, Hidelberg Albuquerque, Márcio Dias, Nádia Da Silva, André Carvalho, Adriano Oliveira, and Carmelo Bastos-Filho. 2024. [Ulysses Tesemô: a new large corpus for Brazilian legal and governmental domain](#). *Language Resources and Evaluation*, 59:1685–1704.
- Ellen Souza, Hidelberg O. Albuquerque, Nádia F. F. Silva, Matheus Cerqueira, Rafael O. Nunes, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2026. [Reconhecimento de Entidades Nomeadas no Domínio Legal: Um Panorama para a Língua Portuguesa](#). In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 4 edition, volume 3, book chapter 13. BPLN.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. [BERTimbau: Pretrained BERT Models for Brazilian Portuguese](#). In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 403–417, Berlin, Heidelberg. Springer-Verlag.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. [brat: a Web-based Tool for NLP-Assisted Text Annotation](#). In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.
- 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.

A Appendix

Verification prompt

Role	You are an expert Brazilian Legal Analyst and Data Verifier specialized in legal citations (referenciação jurídica).
Task	Validate legal identifiers. You must answer a question verifying if the “text” corresponds correctly to the “class” and if the legal reference actually exists in the Brazilian Legal System. You must be extremely accurate, meaning you must check official sources in the internet, such as https://www.planalto.gov.br/, https://www.jusbrasil.com.br/, https://www.al.sp.gov.br/repositorio/legislacao/constituicao/1989/compilacao-constituicao-0-05.10.1989.html, among others.
Verification Variables	• Primary Fields: Use text and class.
Validation Criteria	• Class-Text Alignment: The class must be the most specific level mentioned in the text and must match the existing legal code. • LEGISLACAO Allowlist: This class strictly includes: Constituição, Leis, Decreto-Leis, and Emendas Constitucionais. • Invalid for LEGISLACAO: Súmulas, Jurisprudência, Recomendações, Portarias, Medidas Provisórias (MP), Decretos, Instruções Normativas (IN), and Resoluções. • Factual Existence (Critical): You must verify if the specific subdivision (Article, Paragraph, Inciso) actually exists within the cited Law or Code. • Example: If a text cites “Art. 489, § 11 of the CPC,” and the CPC only goes up to § 3 for that article, it is invalid. • Resolvability: The text must be unique. “CPC” is valid as LEGISLACAO, but “Artigo do CPC” is invalid because the number is missing.
Invalidation Rules	Mark resolved: false if: • The legal subdivision does not exist in the official Brazilian text. • The class is mismatched (e.g., text describes an Inciso but class says ARTIGO). • The item belongs to a forbidden category (e.g., a Portaria). • Vital information (numbers/years) is missing.
Output Format	Return a JSON object, which must contain: - resolved: Boolean (true or false). - reason: A brief, technical explanation in Portuguese (only if resolved is false).
Examples	Example 1: Input: Na legislação brasileira, existe “art . 240 , do CPC”? Esse item jurídico é um “ARTIGO”? Output: {"resolved": True} Example 2: Input: Na legislação brasileira, existe “CPC”? Esse item jurídico é um “LEGISLACAO”? Output: {"resolved": True} Example 3: Input: Na legislação brasileira, existe “art . 240 CPC”? Esse item jurídico é um “ARTIGO”? Output: {"resolved": True} Example 4: Input: Na legislação brasileira, existe “art . 489, § 11, IV, do CPC”? Esse item jurídico é um “INCISO”? Output: {"resolved": False, "reason": "Não existe parágrafo 11 no artigo 489 do CPC"} Example 5: Input: Na legislação brasileira, existe “Código de Processo Civil”? Esse item jurídico é um “LEGISLACAO”? Output: {"resolved": True}
Target Input	Na legislação brasileira, existe “[TARGET ENTITY TEXT]”? Esse item jurídico é um “[TARGET ENTITY CLASS]”?

Figure 6: LLM verification step prompt.

Prompt for direct extraction of legal subcomponents

Task	You are an expert legal data annotator specialized in Brazilian Law. Your task is to perform Named Entity Recognition (NER) on legal texts to identify and classify specific juridical components.
Entity Types	• LEGISLACAO (LAW): Strictly limited to “Constituição”, “Leis”, “Decreto-Leis”, and “Emendas Constitucionais”. • Exclusion: Do not extract “Súmulas”, “Jurisprudência”, “Recomendações”, “Portarias”, “Medidas Provisórias” (MP), “Decretos”, “Instruções Normativas” (IN), or “Resoluções”. • ARTIGO (ART): Direct citations of specific articles. • PARAGRAFO (§): Direct citations of paragraphs. • CAPUT: References to the head/main section of an article. • INCISO (I, II, III...): Citations of Roman numeral items. Also consider “Item” as an “INCISO”. • ALINEA (a, b, c...): Citations of lowercase letter subparagraphs.
Rules	• With the exception to LEGISLACAO, every other entity type MUST be accompanied by all required fragments to properly extract the juridical component. For instance, in the text “... art. 102 da CF ...”, there is “CF” (LEGISLACAO) and “art. 102 da CF” (ARTIGO). Only “art. 102” is a wrong extraction. • Entities can overlap and can be composed by multiple discontinuous fragments. • Extract the entities exactly as they appear in the input text.
Input (example 1) (truncated text)	EMENTA : HABEAS CORPUS - FURTO QUALIFICADO - CONCURSO DE PESSOAS - ... CONDUTAS TIPIFICADAS NO ART . 155 , §4º INCISOS II E IV C/C ART . 288 DO CPB - ... PRESENÇA DOS REQUISITOS DO ART . 312 DO CPP - NÃO HÁ QUE SE FALAR ... Cumprе ressaltar que conforme consta ... nos termos do disposto no art . 312 , p.u do CPP , ... como aconteceu , repetindo , diante de uma decisão fundamentada suficientemente ...
Output (example 1) (truncated list)	1. art . 155 , §4º , II CPB True - INCISO as it is a "INCISO" 2. art . 155 , §4º IV CPB True - INCISO as it is a "INCISO" 3. art . 155 , §4º CPB True - PARAGRAFO as it is a "PARAGRAFO" 4. art . 155 CPB True - ARTIGO as it is a "ARTIGO" 5. ART . 288 DO CPB True - ARTIGO as it is a "ARTIGO" 6. CPB True - LEGISLACAO as it is a "LEGISLACAO" 7. ART . 312 DO CPP True - ARTIGO as it is a "ARTIGO" 8. CPP True - LEGISLACAO as it is a "LEGISLACAO" 9. Decreto 3.048/99 False as it is a "Decreto" [...]
Examples 2-5	[example 2: Input ... / Output ...] [example 3: Input ... / Output ...] [example 4: Input ... / Output ...] [example 5: Input ... / Output ...]
Target Input	[Input Text]

Figure 7: Adapted PromptNER prompt for the task of direct extraction of legal citation subcomponents using unified NER.