Labor Lex: A New Portuguese Corpus and Pipeline for Information Extraction in Brazilian Legal Texts

Pedro Vitor Quinta de Castro[1,2], **Nadia Félix Felipe da Silva**[1]

[1]Universidade Federal de Goiás, [2]Data Lawyer

Correspondence: pedro@datalawyer.com.br

Abstract

Relation Extraction (RE) is a challenging Natural Language Processing task that involves identifying named entities from text and classifying the relationships between them. When applied to a specific domain, the task acquires a new layer of complexity, handling the lexicon and context particular to the domain in question. In this work, this task is applied to the Legal domain, specifically targeting Brazilian Labor Law. Architectures based on Deep Learning, with word representations derived from Transformer Language Models (LM), have shown state-of-the-art performance for the RE task. Recent works on this task handle Named Entity Recognition (NER) and RE either as a single joint model or as a pipelined approach. In this work, we introduce Labor Lex, a newly constructed corpus based on public documents from Brazilian Labor Courts. We also present a pipeline of models trained on it. Different experiments are conducted for each task, comparing supervised training using LMs and In-Context Learning (ICL) with Large Language Models (LLM), and verifying and analyzing the results for each one. For the NER task, the best achieved result was 89.97% F1-Score, and for the RE task, the best result was 82.38% F1-Score. The best results for both tasks were obtained using the supervised training approach.

1 Introduction

Information Extraction (IE) is a field of Natural Language Processing (NLP) that involves a range of tasks aimed at structuring unstructured textual information, thereby facilitating the categorization of such information (Maynard et al., 2016). Among these tasks, two can be considered fundamental to this objective: Named Entity Recognition (NER) (Nadeau and Sekine, 2007) and Relation Extraction (RE) (Zhang et al., 2017). NER aims to identify and classify proper nouns in textual content (Maynard et al., 2016), while RE is responsible for classifying the relationship between a pair of entities. IE tasks, such as NER and RE, are essential for building knowledge bases and graphs (Huang and Wang, 2017), which support and provide inputs for the development of other NLP applications, such as semantic search, summarization, and question-answering (Huang and Wang, 2017).

Formally, the task is defined as a relation classification over entity pairs, represented as the triplet ***(subject, RELATION, object)***, where the ***subject*** and the ***object*** are two entities that share a ***RELATION*** between them. Figure 1 shows the identification of entities, followed by the classification of the relations between them, performed by NER and RE models. In sentence **S1** there are two types of relations between the entities: *grant* corresponds to the *decision* of the judge (**DECISION**) to accept the *unhealthy work conditions allowance* claim (**CLAIM**) made by the plaintiff, while *R$2,230.23* is the *amount* (**CLAIM_VALUE**) that the defendant must pay for such a claim. For entity pairs within a sentence window that do not exhibit any relation, the label assigned is **NO_RELATION**, as in relation **R8** between *compensation for moral damages* and *grant*.

Entity relations can occur across different scopes, depending on the location of the entities involved: (i) the entities can be present in the same sentence (Intra-Sentence); (ii) in different contiguous sentences (Inter-Sentence); or (iii) across all sentences in an entire document. In Figure 1, **R1** and **R2** are examples of Intra-Sentence relations, as both entities in each relation are part of the same sentence **S1**. **R5** is an example of an Inter-Sentence relation, as the subject of the relation (*termination payments*) is in sentence **S3**, and the corresponding decision object (*denied*) is in sentence **S4**. This work contemplates both *Intra* and *Inter-Sentence* relations.

This research addresses the complexities of RE, a challenging task, particularly within the legal domain. The project is motivated by a signifi-

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 184–206

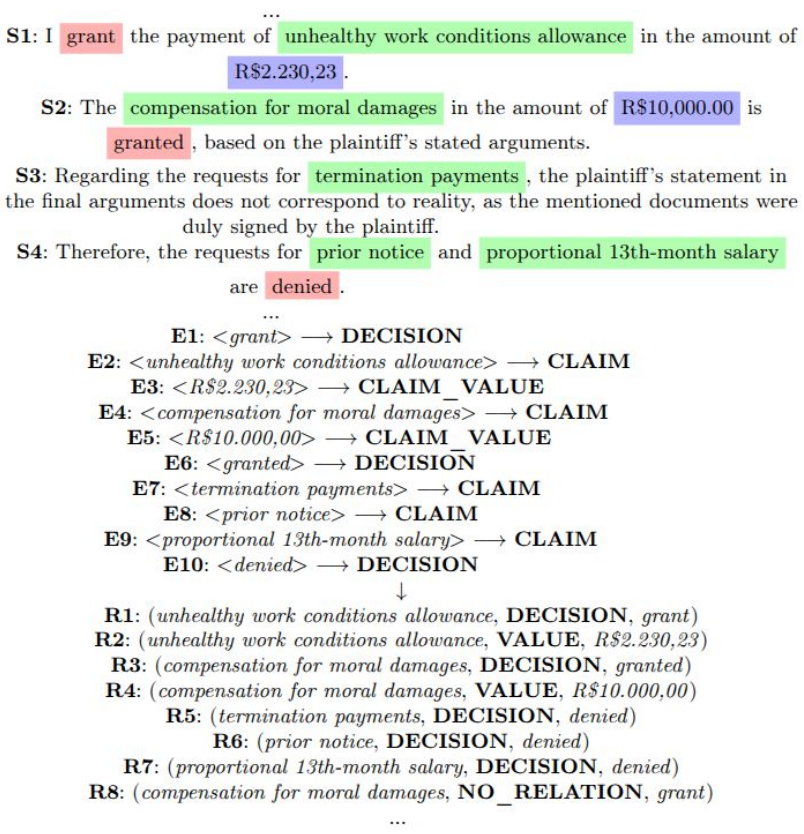

Figure 1: Example of legal entities extracted from a four-sentence window of a labor court judgment, comprising claims, amounts, decisions, and their relation triplets. Only one **NO_RELATION** instance is shown; the others were omitted.

cant gap in resources and benchmarks for the Portuguese language compared to English, which limits the progress of RE in this domain. This domain-specific focus is crucial because general-purpose datasets are not suited for legal-specific information, and the resulting structured data provides valuable insights for ***Jurimetrics***[1] (Jaeger Zabala and Silveira, 2014).

We propose a pipeline-based approach for Relation Extraction, training an NER and RE model for each task using the Labor Lex dataset built for the Portuguese Legal Domain. For the NER task, we evaluate the fine-tuning of different Transformer base models, also evaluating linear and CRF (Conditional Random Field) (Lafferty et al., 2001) classifiers. For RE, we experiment with the PURE architecture (Zhong and Chen, 2021) for a specific domain in the Portuguese language. We also conducted experiments on a joint, end-to-end approach for both tasks, utilizing In-Context Learning (ICL) (Dong et al., 2024) with LLMs and comparing the results obtained with the fine-tuned models. From now on, we refer to this approach as **LLM ICL**.

We present four key contributions: (1) Labor Lex Corpus, a new Portuguese Corpus for the Labor Legal Domain[2]. (2) An Evaluation of NER and RE tasks on this domain dataset. (3) A new NER and RE model pipeline. (4) Results analysis and comparison with the LLM ICL approach.

The remainder of the paper is organized as follows: Section 2 introduces our corpus annotation. Section 3 describes the Related Work. Section 4 presents our proposed Model to train on our data. Section 5 describes the Experimental Evaluation. And finally, Section 6 concludes the paper.

2 Labor Lex Corpus

When initiating a labor lawsuit, the claimant must submit a document known as ***initial petition*** through their representing attorney. The petition must enumerate the ***claims*** and the legal substantiation for each, also identifying the respondent who allegedly breached the labor contract. According to Brazilian Labor Law, the petition must specify the

[1]Statistical analysis applied to legal data.

[2]We intend to release a subset of the full corpus to the research community.

amount claimed[3] for each claim. The sum of all claim values is called the ***case value***. When deciding, the judge may grant or deny each claim made by the claimant. Similarly, even for claims that are granted, the amount awarded may not correspond to the amount claimed. The sum of the amounts granted by the judge for each claim is referred to as the ***conviction value***. At any point in the lawsuit, there can be a settlement between the parties, specifying an agreement that the judge must sanction. The amount agreed upon between the parties, to be paid by the respondent to the claimant, is called the ***settlement value***. There is also the ***legal costs value*** representing the total expenses incurred during a lawsuit, owed to the Judiciary for providing public services.

2.1 Named Entitiy Categories

Other works in the Brazilian Legal Domain have focused on creating corpora for the NER task. Luz de Araujo et al. (2018) introduced ***LeNER-Br***, a NER dataset consisting of 70 legal documents from different justice courts, containing six different types of entities: organizations, persons, time, locations, laws, and precedents. Correia et al. (2022) proposed a corpus for the NER task containing 594 decisions from the Brazilian Supreme Court (STF), focusing on courts, dates, and different types and granularity levels of legal grounds. Albuquerque et al. (2022) proposed ***UlyssesNER-Br***, a corpus composed of 100 bills of law and 500 legislative consultations, in which seven different categories of entities were labeled: legal grounds, organizations, persons, locations, dates, events, and law products. In de Castro (2019), a NER corpus containing 144 documents was created for the Brazilian Labor Legal Domain, focusing on categories such as names of people and organizations, as well as their roles in the lawsuits; values of settlements, cases, convictions, and legal costs. This work introduces new entity categories for the same domain, such as claims, decisions, and claim values, while also introducing annotations for the RE task. The relations defined here aim to associate claims with their respective values and the decisions in judgments that grant or deny such claims. In addition to the claim relations, another relation between roles and people or organizations allows for identifying who the parties in the lawsuits are: their lawyers, witnesses, representatives, experts, and the judge responsible for the rulings. Figure 1 shows different claims made in a lawsuit in terms of their value and decision relations, sampled from a judgment.

We introduce Labor Lex, a novel corpus for the Brazilian Labor Legal Domain, comprising 465 annotated documents. This corpus supports both Named Entity Recognition and Relation Extraction tasks. The entity categories are defined as follows:

1. ***Assignment***: Refers to the entity to whom a specific obligation or ruling in the lawsuit is assigned;

2. ***Claim*** and ***Repercussion***: A *Claim* denotes a legally asserted right or benefit, such as vacation pay. A *Repercussion* signifies the secondary impact of a Claim on other rights. For instance, an overtime Claim may affect the calculation of the 13th salary.

3. ***Decision***: *Decision* is an expression designating a decision in a lawsuit, which can be related to a specific claim or the whole case;

4. ***Organization*** and ***Person***: Names of individuals and organizational entities;

5. ***Role***: Role refers to the function of individuals and organizations within a legal proceeding, such as *claimant*, *defendant*, or *judge*.

6. ***Settlement***, ***Case***, ***Conviction***, ***Legal Costs***, and ***Claim***[4] values: as explained at the beginning of Section 2.

Table 6 in the appendix presents examples of entities of types ***Assignment***, ***Decision***, ***Claim***, and ***Role***, which are the main categories of entities involved in the annotated relations in this work. Examples of ***Repercussion*** are not provided because the mentions are similar to claims, only changing the classification according to the context. The categories ***Person*** and ***Organization*** are also omitted, as they are proper names.

2.2 Relation Categories

The annotations made in this work are *Single Entity Overlap* (SEO) (Wang et al., 2020), meaning each object can be related to multiple subjects: an ***assignment***, a ***decision*** or a ***claim value*** may be related to more than one ***claim*** or ***repercussion***; and a ***role*** may be related to more than one ***person***

[3] Referred to in this work as the ***claim value***.

[4] Amount claimed in petitions or granted in decisions for each claim.

or ***organization***. The following relation categories were annotated:

1. ***Assignment***: This relation resolves possible ambiguities by connectinng ***Claim*** or ***Repercussion*** entities (subjects) to ***Assignment*** entity (object), indicating the responsibility to bear the obligations of a decision concerning that specific claim.

2. ***Decision***: Relation between ***Claim*** or ***Repercussion*** (subjects) and ***Decision*** (object);

3. ***Role***: This relation corresponds to the association between entities of type ***Person*** or ***Organization*** (subjects) and ***Role*** (object). The purpose of this relation is to map the procedural role that each participant has in the case;

4. ***Value***: This relation occurs between entities of type ***Claim*** or ***Repercussion*** (subjects) and ***Claim Value*** (object), indicating the corresponding value of the claim or repercussion.

2.3 Annotation Methodology and Statistics

The annotation tool used was ***INCEpTION*** (Klie et al., 2018)[5]. Two lawyers with previous experience in annotating entities in legal documents annotated the documents. To maximize coverage under fixed annotation resources, documents were partitioned into two non-overlapping subsets and assigned to each of the annotators. Next, a reciprocal cross-review was performed: each annotator reviewed the subset of the other and proposed edits; disagreements were resolved by consensus between the two annotators, yielding a single curated[6] dataset. The review aimed to ensure adherence to annotation criteria and standards, as well as the inherent detection and correction of annotation errors. During the annotation process, the annotators reported that the ***CLAIM*** entity category was the one that triggered the most discussions, as they deemed it the most subjective one. Because the design did not include redundant double annotation of the same documents, the Inter-Annotator Agreement (IAA) was not computed over the full corpus. While this choice prioritized breadth of annotation over duplicated effort, the cross-review and curation procedure served as our quality-control mechanism.

[5] https://inception-project.github.io/

[6] INCEpTION contains a *Curation* feature used for reviewing annotations.

The produced corpus is composed of 465 documents from 149 different cases, distributed among various types: ***Petitions*** (178), ***Contestations*** (56), ***Hearing Records*** (69), ***Judgments*** (90), ***Appeals*** (58), ***Decision*** (9), ***Dispatch*** (1), ***Notification*** (2), and ***Warrant*** (2). Labor Lex has a total number of sentences of 39,905, with 71,146 annotated entities and 15,011 annotated relations. The total number of tokens in the documents is 1,260,965 according to standard whitespace tokenization, and 1,737,904 according to the WordPiece tokenization (Devlin et al., 2019). Table 1 displays the number of annotated entities and relations in each category. Figure 3 in the appendix shows examples of relations annotated in this work.

NER		RE	
Category	**#Entities**	**Category**	**#Relations**
Assignment	1,400	Assignment	2,619
Case Value	535	Decision	5,874
Claim	29,444	Role	4,053
Claim Value	1,829	Value	2,465
Conviction Value	506	**Total**	**15,011**
Court	1,618		
Court Branch	627		
Decision	3,691		
Legal Costs Value	962		
Legal Ground	10,081		
Location	1,985		
Organization	4,735		
Person	4,848		
Proceeding Type	2,438		
Repercussion	2,521		
Role	3,644		
Settlement Value	282		
Total	**71,146**		

Table 1: Entities and Relations annotated for each category.

3 Related Work

Recent works on RE have evolved from traditional neural architectures to Transformer-based models (Vaswani et al., 2017), with advances in both pre-training and task-specific adaptation. ERNIE (Zhang et al., 2019) incorporated structured knowledge into Masked Language Modeling (MLM) pre-training, SpanBERT (Joshi et al., 2020) used span-level objectives, and LUKE (Yamada et al., 2020) employed entity-aware attention. Some works used contrastive learning and masking of entity pairs, such as MTB (Baldini Soares et al., 2019). ERICA (Qin et al., 2021) also uses contrastive learning, but with an objective that focuses on entity and relation discrimination.

Enhancing entity representation through special

markers has proven effective for highlighting entity boundaries, their categories, and roles in the participating relations (Baldini Soares et al., 2019; Peng et al., 2020; Zhong and Chen, 2021; Ye et al., 2022). Yan et al. (2023) do something similar but use an architecture based on Graph Neural Networks (GNN). Joint architectures such as SpERT (Eberts and Ulges, 2019), TPLinker (Wang et al., 2020), and ATLOP (Zhou et al., 2021) combine entity and contextual embeddings for classification.

Graph-based methods like DyGIE (Luan et al., 2019) and DyGIE++ (Wadden et al., 2019) model entities and relations as graph structures. Other works leverage GNN architectures, such as Graph Convolutional Network (GCN) (Zhang et al., 2018) and Attentive Graph Convolutional Network (A-GCN) (Tian et al., 2021), while PL-Marker (Yan et al., 2023) explores various graph topologies. More recently, the Graph Language Model (GLM) by Plenz and Frank (2024) adapts T5 (Raffel et al., 2020) with graph biases to enable joint reasoning over text and graphs.

Prompt-based fine-tuning reformulates RE as a masked prediction task, aligning it with pre-training objectives (Chen et al., 2022; Zhang et al., 2023b; Chen et al., 2024; Efeoglu and Paschke, 2025). More recent work leverages LLMs for zero- and few-shot RE through In-Context Learning (ICL) (Li et al., 2023; Wan et al., 2023; Zhang et al., 2023a; Wang et al., 2023). The RAG4RE framework by Efeoglu and Paschke (2025) extends this by integrating retrieved external knowledge, a key distinction from earlier methods.

For the legal domain, prior work includes a hybrid CRF and rule-based approach for French NER and RE (Andrew, 2018) and a legal triplet extraction system for Chinese (Chen et al., 2020). A recent study by Deußer et al. (2024) applied ICL to seven diverse legal datasets across multiple languages using eleven state-of-the-art LLMs.

In the Brazilian context, research has focused on domain adaptation for various tasks. Polo et al. (2021) adapted word embeddings to classify the status of legal proceedings. More recently, Garcia et al. (2024) introduced Portulex, a benchmark with four datasets for NER and Rhetorical Role Identification. They also performed domain adaptation of RoBERTa (Liu et al., 2019) and evaluated it on this benchmark.

In terms of the available General Domain Portuguese RE corpora, the only benchmark identified was *ReRelEM* (Relations Recognition between Entity Mentions) (Freitas et al., 2008), and few works were found using it (Cardoso, 2008; Bruckschen et al., 2008; Chaves, 2008). Collovini et al. (2020), Reyes et al. (2021), Pavanelli (2022), and da Silva et al. (2023) developed domain-specific corpora targeting commercial and medical applications. There are tasks and evaluations such as those proposed in (Collovini et al., 2019), but since the gold-standard evaluation corpus was not released after the conference, it was not possible to establish a new *benchmark* as a reference for future work. To the best of our knowledge, no prior work addresses Relation Extraction in the Portuguese Legal Domain.

4 Model

We adopt a pipeline where a NER model first identifies entity spans and types, and its outputs are then fed into a RE model. This design is inspired by the PURE framework (Zhong and Chen, 2021), which demonstrated that well-engineered pipeline models can achieve performance comparable to, or even exceeding, that of complex joint models. Figure 2 illustrates the design of our framework.

While the original PURE system employs a span-based NER model as its first stage, our implementation experiments with a Transformer-based encoder combined with two alternative classification layers: a Conditional Random Field (CRF) (Lafferty et al., 2001; Lample et al., 2016; Ma and Hovy, 2016) and a linear neural layer. The Transformer encoder provides rich contextual representations of the input sequence, while the CRF layer offers sequence-level decoding that models label dependencies and enforces valid tag sequences. In contrast, the linear layer performs independent token-level classification, offering a simpler and faster alternative. For the RE stage, we retain the core design principles from PURE, embedding entity spans with special position markers, contextualized via a Transformer encoder, and then classifying them into relation types using the concatenated span representations. This architecture allows the RE model to focus solely on the semantics and context of the provided entities, while enabling a controlled comparison of NER classification strategies.

5 Experimental Evaluation

For our experiments, we trained two models for the NER and RE subtasks. Besides what was described as the proposed model in the previous section, we have also experimented with LLM ICL, providing

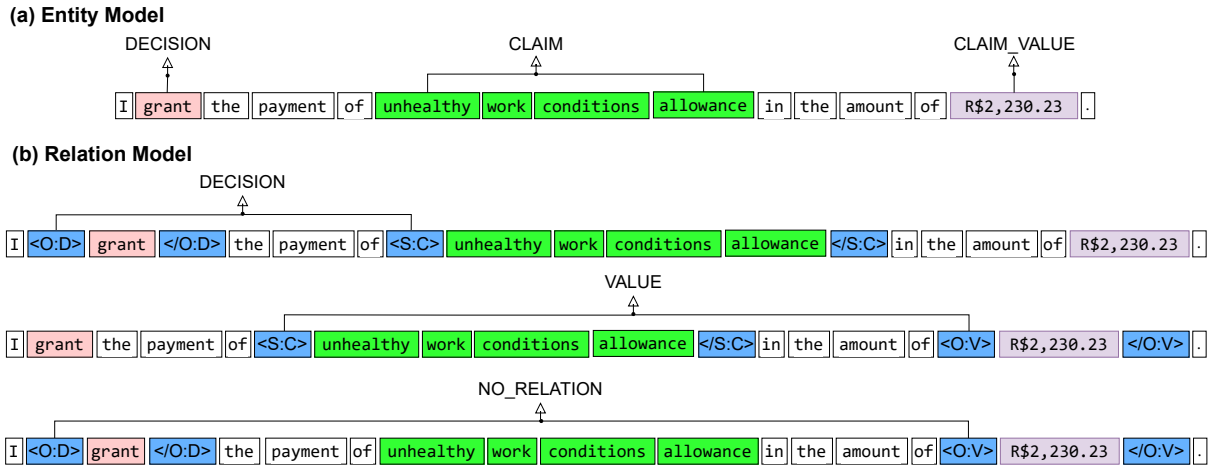

Figure 2: The same S1 sentence example from Figure 1. (a) In this sentence, the expected entities to be extracted by the NER model are “grant” (**DECISION**), "unhealthy work conditions allowance" (**CLAIM**) and “R$2,230.23” (**CLAIM_VALUE**). (b) Next, entity markers are appended around the predicted entities, highlighting them as either subject or object (**S** or **O** prefixes in the markers) of the relation. In the examples, “O:D” indicates that the entity of type **DECISION** is the object of the relation; “S:C” indicates that the entity of type **CLAIM** is the subject of the relation and “O:C” indicates that the entity of type **CLAIM_VALUE** is the object of the relation. The RE model then uses the concatenated contextual embeddings from these markers to classify the relation with a linear layer. As an example, the **CLAIM** entity is related to both the **DECISION** and **CLAIM_VALUE** entities, while unrelated entities are classified as **NO_RELATION**.

LLMs with few-shot examples to evaluate their performance on Labor Lex. Details of the experiments conducted with each approach are presented in the following subsections.

5.1 NER Model

To create the training data for the NER task, we first deduplicated sentences and randomly split them to prevent data leakage from a single legal case into the test set. Specifically, 10% of all sentences were allocated to the test set, while the remaining 90% were used for a 5-fold cross-validation split to create the training and validation sets.

5.1.1 Parameterization, Training and Setup

Three LMs were evaluated for the NER task: the BERT model (Devlin et al., 2019) trained for the Portuguese General Domain by Souza et al. (2020), and two RoBERTa (Liu et al., 2019) models by Garcia et al. (2024) for both General and Legal domains. All evaluated models are of *base* size[7]. The NER task training involves fine-tuning the LM parameters using the NER annotations from Labor Lex. The cost function used to adjust the weights for training with the CRF classifier is the Conditional Log Likelihood, while models using a linear neural layer for classification adopt Cross Entropy.

We evaluated our models using 5-fold cross-validation with three different seeds. This procedure resulted in a total of 15 training runs per model. We evaluated 3 LMs and 2 classifiers, reaching a total of 90 trainings. The reported performance metrics are the F1-Score on the test set, obtained by averaging the 15 training sessions performed for each evaluated combination. The hyperparameters used for these trainings are presented in Table 7 in the appendix section.

5.1.2 Results for NER subtask

Table 2 contains the results for all the experiments conducted with each of the three evaluated language models, considering the two classifier options used. The best classifier was CRF, and the best LM coupled with it was BERT. The best average performance was obtained with the RoBERTaLexPT model Garcia et al. (2024), followed by BERT. With the linear classifier, RoBERTaLexPT outperformed both General Domain models. Appendix Figure 5 contains the evaluation data per entity category for the best NER model. Section A.5.1 in the appendix presents a detailed error analysis for this task.

[7]The base models of BERT and RoBERTa contain approximately 110 million and 125 million parameters, respectively.

Model	LM Domain	F1		Average*
		CRF	Linear	
RoBERTaLexPT (Garcia et al., 2024)	Legal	89.75%	**88.20%**	**88.97%**
BERT (Souza et al., 2020)	General	**89.97%**	87.86%	88.91%
RoBERTaCrawlPT (Garcia et al., 2024)	General	89.58%	87.75%	88.66%

Table 2: Average cross-validation results of NER for each evaluated LM. Results are presented for the CRF and Linear classifiers. The asterisk (*) indicates the overall average while *LM Domain* specifies the domain of LM is pre-trained.

5.2 RE Model

RE training data building is detailed in the Appendix section A.2. The resulting preprocessing of the labeled data produced 7,261 training items (each item is a window of 4 sentences), 675 of which were assigned to the test set, and the remainder was split between the cross-validation sets.

5.2.1 Parameterization, Training and Setup

The three LMs previously evaluated in the NER task were also used for the RE task. Consistent with the NER methodology, the LMs were fine-tuned on the RE annotations from Labor Lex. Cross Entropy is the cost function used to adjust the model weights during training. A 5-fold cross-validation with three different seeds was performed. This resulted in 15 distinct training runs for each evaluated LM (5 folds × 3 seeds), totaling 45 training sessions. Following our pipeline approach, the RE models were trained on gold-standard entities but evaluated on entities predicted by the best-performing NER model. Performance is reported using the F1-Score, averaged across the 15 training runs for each model. Hyperparameters for these trainings are detailed in Table 8 (in appendix).

5.2.2 Results for RE subtask

Table 3 presents the experimental results for these models, including the average for each evaluated LM. The domain-specific LM achieved the best performance in this task, outperforming the next best model by 1.37%. A confusion matrix detailing the results per relation category is presented in Figure 6, and a detailed error analysis is provided in Section A.5.2 (in the appendix section).

5.3 In-Context Learning with LLMs

We evaluated both tasks on our benchmark using different LLMs, applying ICL (LLM ICL) while providing them with few-shot examples to use as references. Two prompting strategies were tested, *Annotation* and *Question Answering*, each one provided with the same 12 few-shot examples (three for each relation type). A structured JSON output is expected for both strategies.

- **Annotation**: This prompting strategy provided the LLMs with details instructions on all entity and relation categories, as well as each property from the JSON data used as input and output. In the prompt, we provide a list of tokens for each example, asking them to fill in the entities and relations data according to the examples given. Appendix A.4.1 contains the prompt used for this strategy.
- **Question Answering**: For this strategy, we provide the LLM with instructions containing the same details regarding the entity and relation categories. However, instead of instructing the LLM to fill in the provided input according to those instructions, we use a series of questions for it to answer in the specified JSON format. Appendix A.4.2 contains the prompt used for this strategy.

For our experiments, we have evaluated the following LLMs: gemini-2.0-flash (Google, 2024), OpenAI o3 (OpenAI, 2025), gpt-4o-mini (OpenAI, 2024), deepseek-chat-v3-0324 (DeepSeek-AI, 2024), gemma-3-27b-it (Team et al., 2025), qwen3-235b-a22b (Team, 2025), and llama-3.1-405b-instruct (Grattafiori et al., 2024). Details on the APIs used for them are presented in appendix section A.4. Each LLM API was called three times to measure the consistency of the results.

5.3.1 Results with LLM ICL

We conducted the experiments using the same scripts and metrics employed in the supervised training approaches, utilizing the same test set produced for each task. We performed post-processing to ensure valid JSON output from the LLMs. Table 4 shows the results grouped by each evaluated LLM and prompt strategy. The best LLM for both tasks was o3 from OpenAI, using both prompt strategies. The best open-weight LLM is deepseek-chat-v3-0324 using QA, surpassing gemini-2.0-flash in the

Model	LM Domain	Precision	Recall	F1
RoBERTaLexPT (Garcia et al., 2024)	Legal	**80.00%**	**85.05%**	**82.38%**
BERT (Souza et al., 2020)	General	78.04%	84.87%	81.01%
RoBERTaCrawlPT (Garcia et al., 2024)	General	77.50%	83.58%	80.23%

Table 3: Average cross-validation results of RE for each evaluated LM. *LM Domain* indicates the domain in which the evaluated LM is pre-trained.

LLM	Prompt Strategy	NER F1	RE F1
deepseek-chat-v3-0324	Annotation	60.42%	41.66%
deepseek-chat-v3-0324	QA	62.26%	46.23%
gemini-2.0-flash	Annotation	66.59%	44.86%
gemini-2.0-flash	QA	65.76%	50.57%
gemma-3-27b-it	Annotation	50.56%	30.37%
gemma-3-27b-it	QA	51.26%	33.28%
gpt-4o-mini	Annotation	40.53%	13.56%
gpt-4o-mini	QA	40.94%	18.16%
llama-3.1-405b-instruct	Annotation	58.46%	38.24%
llama-3.1-405b-instruct	QA	55.77%	37.19%
o3	Annotation	70.46%	56.97%
o3	QA	**71.11%**	**57.98%**
qwen3-235b-a22b	Annotation	53.84%	32.69%
qwen3-235b-a22b	QA	54.12%	35.38%

Table 4: Average results obtained for NER and RE tasks, with each LLM and prompt strategy (Annotation and Question Answering - QA).

NER Model	Approach	F1
Transformers-CRF	BERT	**89.97%**
LLM ICL	o3	71.11%
RE Model	**Approach**	**F1**
PURE	RoBERTaLexPT	**82.38%**
LLM ICL	o3	57.98%

Table 5: Performance comparison for both tasks, between the best supervised approaches of each task and the best performing LLM using ICL.

Annotation approach by 1.37%. Table 10 from the Appendix shows that there is an average improvement of 2.92% of the QA strategy over the Annotation strategy. Appendix Table 9 presents the results grouped by each evaluated LLM.

6 Conclusions

In this paper, we introduced and experimented with Labor Lex, a newly created legal dataset for the NER and RE tasks, evaluating different Transformer-based Language Models, as well as two classifiers for NER, and verifying the best results for the created benchmark with the chosen architecture. The best LM for the NER task was the BERT model from Souza et al. (2020), and the best classifier layer was Conditional Random Fields (Lafferty et al., 2001), reinforcing the results obtained by Lample et al. (2016); Ma and Hovy (2016); Corro et al. (2025), which demonstrated the performance gain from using CRF as the classifier for sequential word classification tasks. For the RE task, the best evaluated LM was RoBERTaLexPT from Garcia et al. (2024), using the PURE framework from Zhong and Chen (2021).

An analysis of the best supervised models indicates that the NER subtask is strong overall - ***CLAIM*** (92.46% F1) and ***CLAIM_VALUE*** (93.65%) - but exhibits lower performance for ***DECISION*** (87.98%) and ***REPERCUSSION*** (91.27%). Figure 5 (In appendix A) reveals confusion between the ***CLAIM*** and ***REPERCUSSION*** categories. The matrix (see Figure 5 of appendix A) also indicates a lower recall for ***CLAIM***, ***CLAIM_VALUE*** and ***DECISION***, missing up to 8.1% of the tokens for the latter category. Boundary mismatches account for 10.49% of NER errors. For the RE subtask, the highest F1-scores are observed for ***ROLE*** (92.41%), followed by ***ASSIGNMENT*** (91.87%) and ***VALUE*** (90.95%), whereas ***DECISION*** attains 81.26% due to a high false positive rate (27.26%). Additional details from the error analysis for both subtasks are presented in the appendix section A.5. These findings indicate that the pipeline reliably links persons/organizations to roles and relates claims to values/assignments, while decision-centric phenomena (at both entity and relation levels) remain the principal bottleneck and the most promising target for future optimization.

Regarding the performance of the LMs, having BERT as the best performing model for the NER task shows some dissonance compared to the results obtained by Garcia et al. (2024) and Liu et al. (2019). While Liu shows improvements by using RoBERTa compared to BERT (Devlin et al., 2019), Garcia demonstrates that the Legal Domain RoBERTaLexPT model has achieved superior performance in benchmarks within the same domain. The NER experiments in this work show that the best results obtained with the CRF classifier were achieved using the BERT model, which outperformed both RoBERTa models from the general and legal domains. A possible motivation for these

results is that the hyperparameter space of the NER task for the RoBERTa models may differ from that of BERT when using the CRF classifier.

For the RE task, the best-performing model was the RoBERTa domain-specific model; however, the BERT-based model still outperformed the General Domain RoBERTaCrawlPT. The experiments for this task provided the best model for the Labor Legal Domain from Brazil, with an F1-Score of 82.38% in the created benchmark.

Our experiments show that supervised training still outperforms few-shot In-Context Learning with LLMs for information extraction tasks, as shown in Table 5. While few-shot ICL offers rapid prototyping without training, it suffers from higher inference costs, latency, and more variable performance. In contrast, supervised models require annotated data and fine-tuning but deliver reliable, low-cost inference once deployed. Overall, supervised methods remain the most cost-effective choice for high-accuracy production use, with few-shot ICL being better suited for quick experimentation or low-resource contexts.

For future work, we plan to extend our experiments on the proposed benchmark by exploring architectural and methodological variations discussed in Section 3, with the goal of further improving task performance. Specifically, we intend to investigate end-to-end joint models for NER and RE, graph-based approaches, and prompt-based fine-tuning strategies. Additionally, we will experiment with data augmentation techniques using LLMs to mitigate data scarcity and enhance model generalization.

Acknowledgments

We gratefully acknowledge the support of the Center of Excellence in Artificial Intelligence at the Federal University of Goiás (CEIA-UFG[8]), whose institutional and computational resources made this work possible. We also thank Data Lawyer[9] for funding this research targeted at the Legal Domain. Finally, we thank our team of annotators for their careful and dedicated work during corpus construction.

References

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 *Computational Processing of the Portuguese Language*, pages 3–14, Cham. Springer International Publishing.

Judith Jeyafreeda Andrew. 2018. Automatic extraction of entities and relation from legal documents. In *Proceedings of the Seventh Named Entities Workshop*, pages 1–8, Melbourne, Australia. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Mırian Bruckschen, José Guilherme Camargo De Souza, Renata Vieira, and Sandro Rigo. 2008. Sistema serelep para o reconhecimento de relaç oes entre entidades mencionadas. *Mota and Santos (Mota and Santos, 2008)*.

Nuno Cardoso. 2008. Rembrandt - reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto. *quot; Encontro do Segundo HAREM (Universidade de Aveiro Portugal 7 de Setembro de 2008)*.

Marcírio Chaves. 2008. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2778–2788, New York, NY, USA. Association for Computing Machinery.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhenbin Chen, Zhixin Li, Yufei Zeng, Canlong Zhang, and Huifang Ma. 2024. Gap: A novel generative context-aware prompt-tuning method for relation extraction. *Expert Systems with Applications*, 248:123478.

[8] https://ceia.ufg.br

[9] https://datalawyer.com.br

Sandra Collovini, Patricia Nunes Gonçalves, Guilherme Cavalheiro, Joaquim Santos, and Renata Vieira. 2020. Relation extraction for competitive intelligence. In *Computational Processing of the Portuguese Language*, pages 249–258, Cham. Springer International Publishing.

Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. 2019. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. In *IberLEF@ SEPLN*, pages 390–410.

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.

Caio Corro, Mathieu Lacroix, and Joseph Le Roux. 2025. Bregman conditional random fields: Sequence labeling with parallelizable inference algorithms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29557–29574, Vienna, Austria. Association for Computational Linguistics.

Diego Pinheiro da Silva, William da Rosa Fröhlich, Blanda Helena de Mello, Renata Vieira, and Sandro José Rigo. 2023. Exploring named entity recognition and relation extraction for ontology and medical records integration. *Informatics in Medicine Unlocked*, 43:101381.

Pedro Vitor Quinta de Castro. 2019. Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico. Available in `http://repositorio.bc.ufg.br/tede/handle/tede/10276`.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Tobias Deußer, Cong Zhao, Lorenz Sparrenberg, Daniel Uedelhoven, Armin Berger, Maren Pielka, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. 2024. A comparative study of large language models for named entity recognition in the legal domain. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4737–4742.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Sefika Efeoglu and Adrian Paschke. 2025. Fine-tuning large language models for relation extraction within a retrieval-augmented generation framework. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 1–7, Vienna, Austria. Association for Computational Linguistics.

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, and Cristina Mota. 2008. Relaçoes semânticas do rerelem: além das entidades no segundo harem. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008*.

Eduardo A. S. Garcia, Nadia F. F. Silva, Felipe Siqueira, Hidelberg O. Albuquerque, Juliana R. S. Gomes, Ellen Souza, and Eliomar A. Lima. 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 Lingustics.

Google. 2024. Introducing gemini 2.0: our new ai model for the agentic era. Model announcement (Google). Introduced as a fast, multimodal model with enhanced reasoning, long context support, and integrated tool usage.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yi Yao Huang and William Yang Wang. 2017. Deep residual learning for weakly-supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1803–1807, Copenhagen, Denmark. Association for Computational Linguistics.

Filipe Jaeger Zabala and Fabiano Feijó Silveira. 2014. Jurimetria: Estatística aplicada ao direito. *Revista Direito e Liberdade*, 16(1):87–103.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and

predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR 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 Computational Processing of the Portuguese Language*, pages 313–323. Springer.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein. 2016. Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 6(2):1–194.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

OpenAI. 2024. Gpt-4o mini: cost-effective intelligence. Model announcement (OpenAI). A small, fast, and affordable multimodal model ("o" for omni) supporting text and image inputs.

OpenAI. 2025. Introducing openai o3 and o4-mini. Model announcement (OpenAI). Advanced reasoning model succeeding the o1 model, designed for math, science, coding and visual perception.

Lucas Aguiar Pavanelli. 2022. *An End-to-End Model for Joint Entity and Relation Extraction in Portuguese*. Ph.D. thesis, PUC-Rio.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Moritz Plenz and Anette Frank. 2024. Graph language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4477–4494, Bangkok, Thailand. Association for Computational Linguistics.

Felipe Polo, Gabriel Mendonça, Kauê Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Ferreira, Leticia Lima, Antônio Maia, and Renato Vicente. 2021. Legalnlp - natural language processing methods for the brazilian legal language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 763–774, Porto Alegre, RS, Brasil. SBC.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. 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 3350–3363, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Daniel De Los Reyes, Douglas Trajano, Isabel Harb Manssour, Renata Vieira, and Rafael H. Bordini. 2021. Entity relation extraction from news articles in portuguese for competitive intelligence based on bert. In *Intelligent Systems*, pages 449–464, Cham. Springer International Publishing.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021. Dependency-driven relation extraction with attentive graph convolutional networks. 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 4458–4471, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147, Singapore. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Zhaohui Yan, Songlin Yang, Wei Liu, and Kewei Tu. 2023. Joint entity and relation extraction with span pruning and hypergraph neural networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7512–7526, Singapore. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Qianqian Zhang, Mengdong Chen, and Lianzhong Liu. 2017. A review on entity relation extraction. In *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pages 178–183.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023a. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Wenjie Zhang, Xiaoning Song, Zhenhua Feng, Tianyang Xu, and Xiaojun Wu. 2023b. Labelprompt: Effective prompt-based learning for relation classification. *Preprint*, arXiv:2302.08068.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Annotation Examples

The Table 6 and Figure 3 provide examples of relations and their participating entities annotated in INCEpTION, along with the Portuguese terms and their English translations.

The specialist annotators considered certain legal documents, such as ***Decisions***, ***Dispatches***, ***Notifications***, and ***Warrants*** , to be less relevant. This is because these documents are more focused on procedural steps rather than containing substantive arguments or decisions (The ***Decision*** document type is usually used more for communicating that a decision has been taken rather than the actual decision), resulting in a lower volume of useful information for the analysis. Consequently, these document types were underrepresented in the dataset.

A.2 Training Hyperparameters

The Tables 7 and 8 present the hyperparameters used for training the models for NER and RE, respectively.

Hyperparameter	Value
Batch size	8
Learning rate	3×10^{-5}
Learning rate (CRF layer)	7.5×10^{-3}
Gradient accumulation steps	4
Warmup ratio	10%
Weight decay	0.1
Dropout	0.2
Epochs	10

Table 7: Hypeparameters used for NER model training.

Hyperparameter	Value
Batch size	20
Learning rate	2×10^{-5}
Warmup ratio	10%
Weight decay	0.01
Epochs	5

Table 8: Hyperparameters used for RE model training.

Relation Dataset Preprocessing

The following preprocessing and split were done before the RE training model: **(i)** Windows of sentences were created with a size of 4 and a stride of 2[10]. This process is illustrated in Figure 4; **(ii)** Windows of sentences with no inner relations are discarded; **(iii)** The sentences in the windows are joined using the [unused99] token from the Transformer vocabulary[11]; **(iv)** Relations containing at least one entity outside the window are discarded[12]; **(v)** Deduplication of windows of sentences; **(vi)** Random distribution of the windows from the 149 labeled cases for 5-Fold cross-validation, keeping all windows from the same case in the same dataset.

The windows of sentences were created with a size of 4 and a stride of 2 to reflect a real situation when inferring the model on new documents, as it is unknown where the entities participating in relations will be found. Not using a stride (stride = 0) would imply no overlap at all, causing relations that contain entities in different windows to be missed; a greater number of relations would be discarded in Step 3. The cross-validation distribution at the case level is designed to prevent data leakage among datasets, as the overlap of sentences results in the same sentences being present in different windows.

A.3 Evaluation Details

The Figures 5 and 6 are the confusion matrices for the best NER and RE models trained for this work, respectively.

A.4 LLM ICL Experiments Details

For the Gemini model, we used the Vertex AI API [13]; for the OpenAI models, we used their own API [14] as well; and for all other models, we used the API provided by OpenRouter [15]. Regarding the post-processing of the LLMs response, we followed (Deußer et al., 2024) and developed a code-based solution to map the positions of the extracted entities returned from the LLM to the provided list of tokens as input.

[10] Size 4 was chosen due to 1.48% of labeled relations have their entities separated by more than 3 sentences. A stride of 2 was utilized to produce a 2-sentence overlap between adjacent windows.

[11] This is an unused, reserved slot in the tokenizer vocabulary.

[12] In such cases, it is possible that the object or subject of the relation is in a sentence outside the window.

[13] https://cloud.google.com/vertex-ai

[14] https://openai.com/api

[15] https://openrouter.ai

Assignment	Decision
1ª RECLAMADA (1st defendant)	A SUCUMBÊNCIA (The award of legal costs)
A SEGUNDA RECLAMADA (The second defendant)	ACOLHO EM PARTE (I partially grant)
Município demandado (Defendant Municipality)	ACORDO HOMOLOGADO (Approved settlement)
O executado (The executed party)	APELO PROVIDO (Appeal granted)
PARTE AUTORA (Plaintiff)	Arquivem-se definitivamente os autos (Let the case records be definitively archived)
UNIÃO FEDERAL (FEDERAL GOVERNMENT)	Condenar (To condemn)
a litisconsorte (the joint defendant)	DAR PROVIMENTO EM PARTE (To partially grant the appeal)
ao Sindicato (to the Union)	DECIDO CONHECER (I decide to take cognizance)
parte requerente (petitioner)	Defiro (I grant)
impetrante (petitioner)	JULGAR PROCEDENTES (To rule in favor)
Claim	**Role**
01-SALDO DE SALÁRIO (01-Wage balance)	advogado (lawyer)
03-MULTA ART . 467 DA CLT (03-FINE ART. 467 OF THE CLT)	Desembargador Relator (Reporting Justice)
1 / 3 Constitucional de Férias (1/3 Constitutional Vacation Pay)	EXEQÜENTE (Enforcing party/claimant)
13º salário proporcional (Proportional 13th-month salary)	Juiz (Judge)
16hs extras semanais (16 weekly overtime hours)	Julgador de Primeiro Grau (First-degree judge)
40 % de multa sobre o FGTS (40% fine on FGTS)	Juíza do Trabalho Substituta (Substitute Labor Judge)
APLICABILIDADE DA REFORMA TRABALHISTA (APPLICABILITY OF THE LABOR REFORM)	preposto do (a) reclamado (a) (representative of the defendant)
ILEGITIMIDADE ATIVA (LACK OF STANDING TO SUE)	RECLAMADO (DEFENDANT)
ASSISTENCIA JUDICIÁRIA GRATUITA (FREE LEGAL ASSISTANCE)	Relatora Ministra (Reporting Minister)
HONORÁRIOS ADVOCATÍCIOS E SUCUMBENCIAIS (ATTORNEY AND LEGAL COSTS FEES)	Rel . Des . (Reporting Justice)

Table 6: Examples of assignments, decisions, claims and roles entities, which are the main categories of entities participating in the relations annotated in this work. Each example in Portuguese includes its respective translation to English in parentheses.

Figure 3: Examples of relations and their participating entities annotated in INCEpTION. These examples were translated from Portuguese.

Concerning the prompts, we translated them into English for presentation in this work, also adding some translations in parentheses for the actual Portuguese terms corresponding to entities and relation types. Figure 7 contains a sample of one of the JSON objects presented to the LLMs to be used as few-shot examples.

A.4.1 Prompt for Annotation Approach

I will now present a list of samples containing texts, along with the corresponding lists of entities and relations between the entities extracted from these texts.

1. The entities are of the following types:
   (a) "ATRIBUICAO" (ASSIGNMENT): corresponds to the entity category that represents the party being assigned or entrusted with a certain obligation or sentencing in the case. Examples: "à reclamada" (to the defendant), "o executado" (the executed party), and "parte autora" (plaintiff).
   (b) "DECISAO" (DECISION): expresses a decision being made in a judgment, which may be related to a claim or to the case as a whole. Examples: "condenar" (to convict), "deferir" (to grant), "acolher" (to accept), "dar provimento" (to uphold), and "conhecer" (to hear a case).

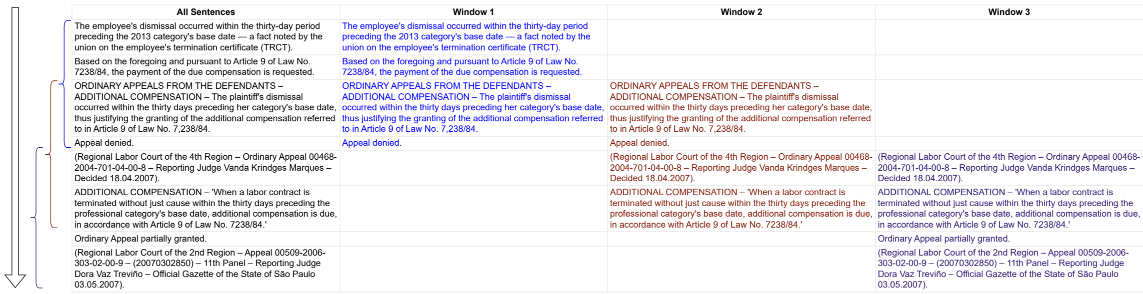

Figure 4: Example of the preprocessing for creating windows of sentences of size 4 for training the RE model.

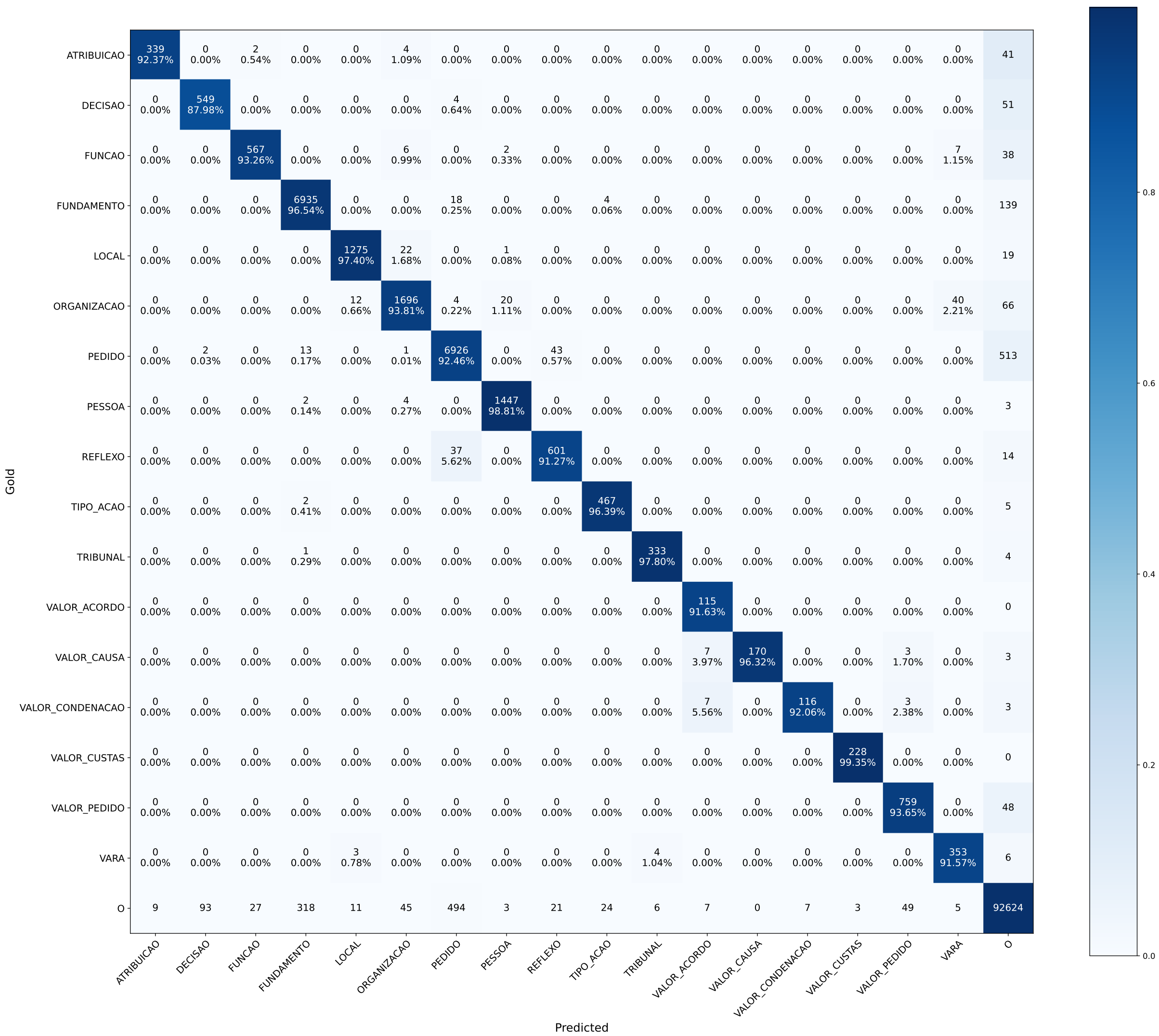

Figure 5: Confusion matrix for the best NER model trained in the experiments conducted. The translated categories from Portuguese to English are ATRIBUICAO: ASSIGNMENT; DECISAO: DECISION; FUNCAO: ROLE; FUNDAMENTO: LEGAL_GROUND; LOCAL: LOCATION; ORGANIZACAO: ORGANIZATION; PEDIDO: CLAIM; PESSOA: PERSON; REFLEXO: REPERCUSSION; TIPO_ACAO: PROCEEDING_TYPE; TRIBUNAL: COURT; VALOR_ACORDO: SETTLEMENT_VALUE; VALOR_CAUSA: CASE_VALUE; VALOR_CONDENACAO: CONVICTION_VALUE; VALOR_CUSTAS: LEGAL_COSTS_VALUE; VALOR_PEDIDO: CLAIM_VALUE; VARA: COURT_BRANCH.

(c) "FUNCAO" (ROLE): corresponds to the function or role of the people mentioned in the documents. Functions are only identified if they accom-

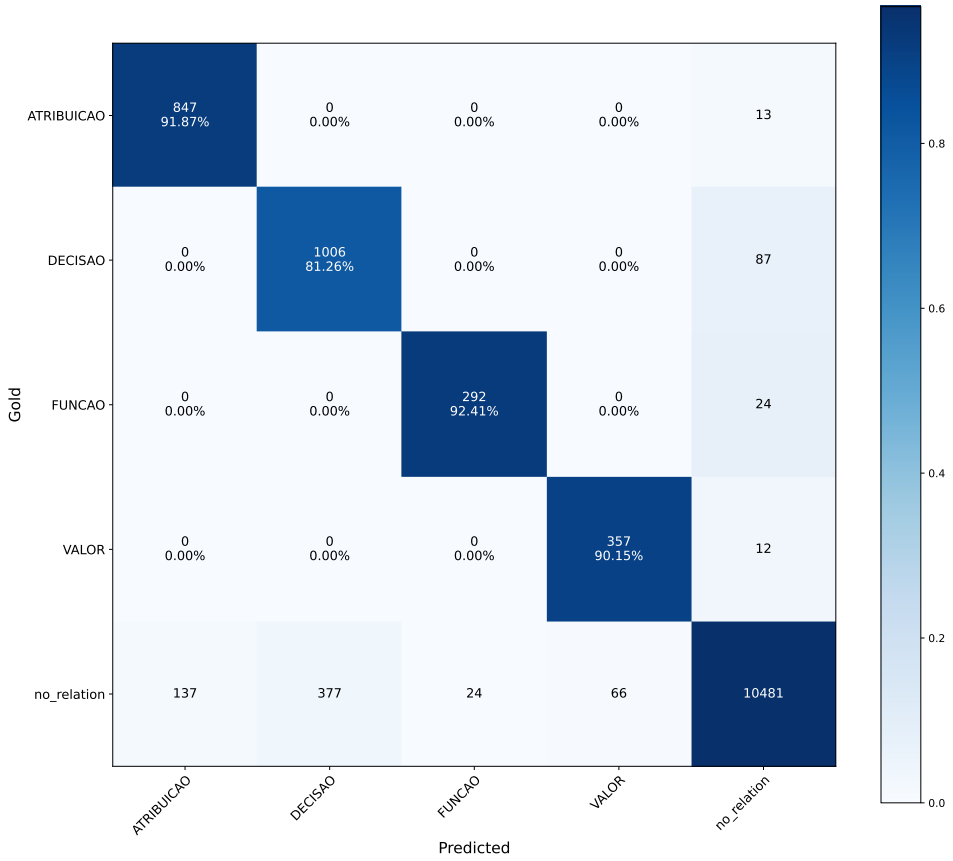

Figure 6: Confusion matrix for the best RE model trained in the experiments conducted. The translated categories from Portuguese to English are ATRIBUICAO: ASSIGNMENT; DECISAO: DECISION; FUNCAO: ROLE; VALOR: VALUE.

pany a "PESSOA" or "ORGANIZACAO". Examples: "advogado" (lawyer), "juiz" (judge), "preposto" (representative), "testemunha" (witness), "reclamante" (claimant), and "reclamado" (defendant).

(d) "FUNDAMENTO" (LEGAL_GROUND): category assigned to any legal provision that may be referenced in the documents to support the claims from lawyers and decisions from judges. Examples: "art. 12, II, do CPC" (Art. 12, II, of the Civil Procedure Code); "artigo 114, inciso VIII, da Constituição Federal" (Article 114, item VIII, of the Federal Constitution); "EMENDA CONSTITUCIONAL Nº 45/04" (Constitutional Amendment No. 45/04).

(e) "LOCAL" (LOCATION): proper names that identify streets, neighborhoods, cities, states, and addresses.

(f) "ORGANIZACAO" (ORGANIZATION): proper names that identify legal entities, which may be companies, institutions, government agencies, associations, foundations, etc. Examples: "Banco Itaú" (Itaú Bank), "Estado do Ceará" (State of Ceará), "INSS", "Justiça do Trabalho" (Labor Court), and "Receita Federal" (Federal Revenue Service).

(g) "PEDIDO" (CLAIM): In the context of Brazilian Labor Justice, claims are formal requests made by the employee (claimant) to the judge, seeking that the company (defendant) be sentenced to fulfill obligations or pay amounts arising from the employment relationship. They define the scope of the action and must be clear, specific, and based on facts and rights. Examples: "Saldo de salário" (salary balance), "aviso prévio" (prior notice), "13º salário proporcional" (proportional 13th salary), "férias pro-

porcionais + 1/3" (proportional vacation + 1/3), "multa de 40% do FGTS" (40% FGTS fine), "horas extras" (overtime hours), "adicional noturno" (night shift bonus), "adicional de insalubridade" (hazard pay), "diferença salarial" (salary difference), "equiparação salarial" (salary equalization), "indenizações por danos morais" (moral damages compensation), and "reconhecimento de vínculo empregatício" (employment relationship recognition).

(h) "PESSOA" (PERSON): proper names that identify natural persons, either full or partial names.

(i) "REFLEXO" (REPERCUSSION): In labor law, "reflexos trabalhistas" (labor repercussions) refer to the financial impacts or consequences that the recognition or payment of a certain main compensation has on other salary or severance payments. It is essentially the cascading effect that one amount has on the calculation of others due to its remunerative nature. Examples: "13º salário" (13th salary), "férias" (vacation), and "FGTS" (service time compensation).

(j) "TIPO_ACAO" (PROCEEDING_TYPE): Types of legal actions correspond to the activities carried out by judges and courts during procedural steps. Examples: "recurso de revista" (appeal for review), "embargos de declaração" (motion for clarification), "apelação cível" (civil appeal), "contrarrazões" (counterarguments).

(k) "TRIBUNAL" (COURT): specific category of organizations that, in the legal context, identify court names. Examples: "STF" (Federal Supreme Court), "STJ" (Superior Court of Justice), "TJMG" (Court of Justice of Minas Gerais), "Tribunal Regional do Trabalho da 21ª Região" (Regional Labor Court of the 21st Region), "Tribunal de Justiça de Goiás" (Court of Justice of Goiás).

(l) "VARA" (COURT_BRANCH): specific category of organizations that, in the legal context, identify labor or judicial court branches. Examples: "11ª Vara do Trabalho de Recife" (11th Labor Court of Recife), "TERCEIRA VARA DO TRABALHO DE MOSSORÓ / RN" (Third Labor Court of Mossoró/RN).

(m) "VALOR_ACORDO" (SETTLEMENT_VALUE): monetary amount related to a settlement being ratified or declared in the document, which the employer agrees to pay to the employee.

(n) "VALOR_CAUSA" (CASE_VALUE): monetary amount corresponding to the total requested by the claimant employee, defined as the claim amount for the case.

(o) "VALOR_CONDENACAO" (CONVICTION_VALUE): monetary amount corresponding to the total sentence set by a judge in a statement indicating the decision in the case, defined as the amount of the sentence to be paid.

(p) "VALOR_CUSTAS" (LEGAL_COSTS_VALUE): monetary amount corresponding to court costs, which are fees paid to the court to cover the costs of the process.

(q) "VALOR_PEDIDO" (CLAIM_VALUE): refers to the identification and quantification of the amounts requested. These are the exact financial amounts the employee seeks to receive, including the main value, its effects on other payments, monetary correction, and applicable interest. Examples are monetary values, either prefixed or not with the currency symbol.

(r) The difference between the "valor de condenação" (conviction value) and the "valor por pedido" (claim value) in a case is that the conviction value refers to the total value of claims granted by the judge. In contrast, the claim value may be used both in a claim or in the conviction context, for one or more specific claims.

(s) The words "pedido" (claim), "reflexo" (repercussion), "organização" (organization), and "pessoa" (person) must not be part of entities; do not include them in the list of identified entities.

(t) In general, the name of each category should not be considered an entity.

(u) Entities should also not start with definite articles such as "a" (feminine singular),

"o" (masculine singular), "as" (feminine plural), and "os" (masculine plural).

2. The relations are of the following types. The rules regarding which entities can be subjects or objects of the relations are:
   (a) "VALOR" (VALUE): only entities of type "PEDIDO" or "REFLEXO" (subjects) can relate to entities of type "VALOR_PEDIDO" (objects).
   (b) "DECISAO" (DECISION): only entities of type "PEDIDO" or "REFLEXO" (subjects) can relate to entities of type "DECISAO" (objects).
   (c) "ATRIBUICAO" (ASSIGNMENT): only entities of type "PEDIDO" or "REFLEXO" (subjects) can relate to entities of type "ATRIBUICAO" (objects).
   (d) "FUNCAO" (ROLE): only entities of type "PESSOA" or "ORGANIZACAO" (subjects) can relate to entities of type "FUNCAO" (objects).
   (e) No relation allows the participation of two entities of the same type, or types different from those described above.

3. You will receive a list of texts to be used for identifying all the types of entities and relations above, in JSON format: Input = "id": int, "tokens": List[str].

4. Entities must be provided in the JSON format Entity = "type": str, "text": List[str], "id": int, where:
   (a) "type": entity type, from the list in step 1;
   (b) "text": list of tokens of the entity;
   (c) "id": sequential numeric ID in the list of entities already provided;
   (d) The tokens in the "text" field must be identical to those in the "tokens" list from the input.

5. Relations must be provided in the JSON format Relation = "type": str, "head": Entity, "tail": Entity, where:
   (a) "type": relation type, from the list in step 2;
   (b) "head": subject of the relation, the entity that initiates or performs the relation;
   (c) "tail": object of the relation, the entity that receives or is affected by the relation.

6. The output for each sample must be in the JSON format Output = "id": int, "entities": list[Entity], "relations": list[Relation].
   (a) The output must be obtained by adding the identified entities and relations, keeping the same numeric ID from the input.

7. The response must not contain comments or markdown, only the JSON output.

8. Create exactly one response for each sample in the input list.

A.4.2 Prompt for Question Answering Approach

I will now present a list of samples containing texts, along with the corresponding lists of entities and relations between the entities extracted from these texts.

1. The entities are of the following types:
   (a) "ATRIBUICAO" (ASSIGNMENT): corresponds to the entity category that represents the party being assigned or entrusted with a certain obligation or sentencing in the case. Examples: "à reclamada" (to the defendant), "o executado" (the executed party), and "parte autora" (plaintiff).
   (b) "DECISAO" (DECISION): expresses a decision being made in a judgment, which may be related to a claim or to the case as a whole. Examples: "condenar" (to convict), "deferir" (to grant), "acolher" (to accept), "dar provimento" (to uphold), and "conhecer" (to hear a case).
   (c) "FUNCAO" (ROLE): corresponds to the function or role of the people mentioned in the documents. Functions are only identified if they accompany a "PESSOA" or "ORGANIZACAO". Examples: "advogado" (lawyer), "juiz" (judge), "preposto" (representative), "testemunha" (witness), "reclamante" (claimant), and "reclamado" (defendant).
   (d) "FUNDAMENTO" (LEGAL_GROUND): category assigned

to any legal provision that may be referenced in the documents to support the claims from lawyers and decisions from judges. Examples: "art. 12, II, do CPC" (Art. 12, II, of the Civil Procedure Code); "artigo 114, inciso VIII, da Constituição Federal" (Article 114, item VIII, of the Federal Constitution); "EMENDA CONSTITUCIONAL Nº 45/04" (Constitutional Amendment No. 45/04).

(e) "LOCAL" (LOCATION): proper names that identify streets, neighborhoods, cities, states, and addresses.

(f) "ORGANIZACAO" (ORGANIZATION): proper names that identify legal entities, which may be companies, institutions, government agencies, associations, foundations, etc. Examples: "Banco Itaú" (Itaú Bank), "Estado do Ceará" (State of Ceará), "INSS", "Justiça do Trabalho" (Labor Court), and "Receita Federal" (Federal Revenue Service).

(g) "PEDIDO" (CLAIM): In the context of Brazilian Labor Justice, claims are formal requests made by the employee (claimant) to the judge, seeking that the company (defendant) be sentenced to fulfill obligations or pay amounts arising from the employment relationship. They define the scope of the action and must be clear, specific, and based on facts and rights. Examples: "Saldo de salário" (salary balance), "aviso prévio" (prior notice), "13º salário proporcional" (proportional 13th salary), "férias proporcionais + 1/3" (proportional vacation + 1/3), "multa de 40% do FGTS" (40% FGTS fine), "horas extras" (overtime hours), "adicional noturno" (night shift bonus), "adicional de insalubridade" (hazard pay), "diferença salarial" (salary difference), "equiparação salarial" (salary equalization), "indenizações por danos morais" (moral damages compensation), and "reconhecimento de vínculo empregatício" (employment relationship recognition).

(h) "PESSOA" (PERSON): proper names that identify natural persons, either full or partial names.

(i) "REFLEXO" (REPERCUSSION): In labor law, "reflexos trabalhistas" (labor repercussions) refer to the financial impacts or consequences that the recognition or payment of a certain main compensation has on other salary or severance payments. It is essentially the cascading effect that one amount has on the calculation of others due to its remunerative nature. Examples: "13º salário" (13th salary), "férias" (vacation), and "FGTS" (service time compensation).

(j) "TIPO_ACAO" (PROCEEDING_TYPE): Types of legal actions correspond to the activities carried out by judges and courts during procedural steps. Examples: "recurso de revista" (appeal for review), "embargos de declaração" (motion for clarification), "apelação cível" (civil appeal), "contrarrazões" (counterarguments).

(k) "TRIBUNAL" (COURT): specific category of organizations that, in the legal context, identify court names. Examples: "STF" (Federal Supreme Court), "STJ" (Superior Court of Justice), "TJMG" (Court of Justice of Minas Gerais), "Tribunal Regional do Trabalho da 21ª Região" (Regional Labor Court of the 21st Region), "Tribunal de Justiça de Goiás" (Court of Justice of Goiás).

(l) "VARA" (COURT_BRANCH): specific category of organizations that, in the legal context, identify labor or judicial court branches. Examples: "11ª Vara do Trabalho de Recife" (11th Labor Court of Recife), "TERCEIRA VARA DO TRABALHO DE MOSSORÓ / RN" (Third Labor Court of Mossoró/RN).

(m) "VALOR_ACORDO" (SETTLEMENT_VALUE): monetary amount related to a settlement being ratified or declared in the document, which the employer agrees to pay to the employee.

(n) "VALOR_CAUSA" (CASE_VALUE): monetary amount corresponding to the total requested by the claimant employee, defined as the claim amount for the case.

(o) "VALOR_CONDENACAO" (CONVICTION_VALUE): monetary amount cor-

responding to the total sentence set by a judge in a statement indicating the decision in the case, defined as the amount of the sentence to be paid.

(p) "VALOR_CUSTAS" (LEGAL_COSTS_VALUE): monetary amount corresponding to court costs, which are fees paid to the court to cover the costs of the process.

(q) "VALOR_PEDIDO" (CLAIM_VALUE): refers to the identification and quantification of the amounts requested. These are the exact financial amounts the employee seeks to receive, including the main value, its effects on other payments, monetary correction, and applicable interest. Examples are monetary values, either prefixed or not with the currency symbol.

(r) The difference between the "valor de condenação" (conviction value) and the "valor por pedido" (claim value) in a case is that the conviction value refers to the total value of claims granted by the judge. In contrast, the claim value may be used both in a claim or in the conviction context, for one or more specific claims.

(s) The words "pedido" (claim), "reflexo" (repercussion), "organização" (organization), and "pessoa" (person) must not be part of entities; do not include them in the list of identified entities.

(t) In general, the name of each category should not be considered an entity.

(u) Entities should also not start with definite articles such as "a" (feminine singular), "o" (masculine singular), "as" (feminine plural), and "os" (masculine plural).

2. The relations are of the following types. The rules regarding which entities can be subjects or objects of the relations are:

   (a) "VALOR" (VALUE): only entities of type "PEDIDO" or "REFLEXO" (subjects) can relate to entities of type "VALOR_PEDIDO" (objects).

   (b) "DECISAO" (DECISION): only entities of type "PEDIDO" or "REFLEXO" (subjects) can relate to entities of type "DECISAO" (objects).

   (c) "ATRIBUICAO" (ASSIGNMENT): only entities of type "PEDIDO" or "REFLEXO" (subjects) can relate to entities of type "ATRIBUICAO" (objects).

   (d) "FUNCAO" (ROLE): only entities of type "PESSOA" or "ORGANIZACAO" (subjects) can relate to entities of type "FUNCAO" (objects).

   (e) No relation allows the participation of two entities of the same type, or types different from those described above.

3. You will receive a list of text samples to be used to answer questions aimed at identifying all types of entities and relations above.

   (a) The samples must be in the format JSON Input = "id": int, "tokens": List[str].

4. Given the entity types listed above in step 1, search for all occurrences of each type in each provided sample.

5. The responses referring to entities must be in the format JSON Entity = "type": str, "text": List[str], "id": int, where:

   (a) "type": entity type, from the list in step 1;

   (b) "text": list of tokens of the entity;

   (c) "id": sequential numeric ID in the list of entities already provided;

   (d) The tokens in the "text" field must be identical to those in the "tokens" list from the input.

6. The relations must be returned in the format JSON Relation = "type": str, "head": Entity, "tail": Entity, where:

   (a) "type": relation type, from the list in step 2;

   (b) "head": subject of the relation, the entity that initiates or performs the relation;

   (c) "tail": object of the relation, the entity that receives or is affected by the relation.

7. Answer the following questions as a list of entities, as described in step 5:

   (a) Which procedural claims and their repercussions are identified in the text?

   (b) Which assignments are identified in the text?

(c) Which decisions are identified in the text?

(d) Which procedural roles of people are mentioned in the text?

(e) What are the claim and repercussion values identified in the text?

8. Answer the following questions as a list of relations, as described in step 6:

   (a) What are the values of each claim and repercussion identified in the text? Link the claim value entities to their respective claims or repercussions in the form of relations.

   (b) To whom was each claim and repercussion identified in the text assigned? Link the assignment entities to their respective claims or repercussions in the form of relations.

   (c) How was each claim or repercussion decided? Link the decision entities to their respective claims or repercussions in the form of relations.

   (d) What is the role of each person or organization identified in the text? Link the role entities to their respective people or organizations in the form of relations.

9. The output for each sample must be in the format JSON Output = "id": int, "entities": List[Entity], "relations": List[Relation].

   (a) The output must be obtained by adding the identified entities and relations, keeping the same numeric ID from the input.

10. The answer must not contain comments or markdown, only the output JSON.

11. Create exactly one answer for each sample in the input list.

12. All entities and relations identified from the samples must be consolidated into a single list, according to the format specified in step 9.

A.4.3 JSON Examples

We used a total of the same 12 few-shot examples for both prompting approaches. Here, we present one example.

LLM	NER F1	RE F1
deepseek-chat-v3-0324	61.34%	43.94%
gemini-2.0-flash	66.18%	47.71%
gemma-3-27b-it	50.91%	31.82%
gpt-4o-mini	40.74%	15.86%
llama-3.1-405b-instruct	57.11%	37.71%
o3	70.78%	57.48%
qwen3-235b-a22b	53.98%	34.04%

Table 9: Average results obtained for each evaluated LLM.

Prompt Strategy	NER F1	RE F1
QA	57.32%	39.83%
Annotation	57.27%	36.91%

Table 10: Average results obtained for each evaluated prompt approach.

A.4.4 Additional Results

The Tables 9 and 10 display additional results for the LLM experiments for both tasks. Table 9 groups the results by LLM model for each task, and Table 10 groups the results by prompt strategy.

A.5 Error Analysis

We selected the best supervised model from each task to conduct an error analysis.

A.5.1 NER model

For the best NER model, according to the confusion matrix presented in Figure 5, we verified that the lowest F1 scores per category were ***DECISION*** (87.98%) and ***REPERCUSSION*** (91.27%). The ***CLAIM*** category had an F1 of 92.46%, and ***CLAIM_VALUE*** had an F1 of 93.65%. Most of the errors concerning repercussions involve predicting them as claims, which occurred for 5.62% of the tokens. The same can be shown for claims; they are most confounded for repercussions as well, for 0.57% of the tokens. However, most of the errors for claims are in terms of recall, with the model missing 6.8% of the tokens labeled as claims. For the decision entities, 0.64% of the tokens were predicted as claims, and 8.1% of the tokens were missed by the model. For the claim values, 5.9% of the tokens were missed by the model. 10.49% of the errors from the model are related to missing the boundaries of the entities, meaning that entities predicted by the model had additional tokens or missed some that were part of the annotations.

Table 11 contains examples of errors for the

```
input: {
    "id": 1,
   "tokens": ['Improcedem', 'os', 'demais', '.', '[unused99]', 'Tudo', 'nos',
   'termos', 'da', 'fundamentação', ',', 'que', 'integra', 'este', 'dispositivo',
   '.', '[unused99]', 'Juros', 'e', 'atualização', 'monetária', 'na', 'forma',
    'da', 'lei', ',', 'observados', 'os', 'parâmetros', 'contidos', 'na',
    'fundamentação', '.', '[unused99]', 'Custas', 'de', 'R', '$', '100', ',',
    '00', ',', 'calculadas', 'sobre', '5', '.', '000', ',', '00', ',', 'ora',
   'arbitrado', 'à', 'condenação', '-', 'art', '.', '789', ',', '§', '2º', ',',
    'CLT', '-', ',', 'pela', 'reclamada', '.']
}
output: {
    "id": 1,
    "entities": [
    {"type": 'DECISAO', "text": ['Improcedem'], "id": 0},
    {"type": 'PEDIDO', "text": ['Juros'], "id": 1},
    {"type": 'PEDIDO', "text": ['atualização', 'monetária'], "id": 2},
    {"type": 'PEDIDO', "text": ['Custas'], "id": 3},
    {"type": 'VALOR_CUSTAS', "text": ['R', '$', '100', ',', '00'], "id": 4},
   {"type": 'VALOR_CONDENACAO', "text": ['5', '.', '000', ',', '00'], "id": 5},
   {"type": 'FUNDAMENTO', "text": ['art', '.', '789', ',', '§', '2º', ',', 'CLT'],
    "id": 6},
    {"type": 'ATRIBUICAO', "text": ['pela', 'reclamada'], "id": 7}
    ],
    "relations": [
    {
        "type": 'ATRIBUICAO',
        "head": {"type": 'PEDIDO', "text": ['Custas'], "id": 3},
      "tail": {"type": 'ATRIBUICAO', "text": ['pela', 'reclamada'], "id": 4}
    }
    ]
}
... Total of 12 examples ...
```

Figure 7: Examples used for both Prompt approaches.

NER task. Tokens in green are tokens from entities that have been correctly classified. Tokens in red are tokens from entities that were incorrectly classified. For the "Gold Entities" column, categories in parentheses next to the tokens of the entities display the annotated label. For the "Predicted Entities" column, the category in parentheses corresponds to the predicted label.

A.5.2 RE model

The best performing category, according to Figure 6, for the RE task was ***ROLE***, with 92.41%, followed by ***ASSIGNMENT*** with 91.87%, ***VALUE*** with 90.95%, and lastly, ***DECISION***, which was the only one below 90%, at 81.26%. The confusion matrix shows that the only category for which the recall is higher than the precision is ***ROLE***, missing 7.6% of the annotations out of a total of 316 in the test set. For the decision category, 27.26% of the predictions made by the model were false positives, leading to the lowest F1 score among the four categories. These results show that the model performs well in associating people and organization names with their roles in the documents. The RE model struggles more with associating claims with their decisions but shows better performance in relating them to claim values and assignments. Table 12 contains examples of ***DECISION*** errors for this task.

Sentences	Gold Entities	Predicted Entities
Therefore, the payment of overtime due, as well as the return to the job, considering the period of stability not respected by the defendant, or related compensation, should also reflect on their severance pay the overtime hours worked during the period.	overtime due (CLAIM) period of stability (CLAIM) compensation (CLAIM) severance pay (CLAIM) overtime hours (CLAIM)	overtime due (CLAIM) period of stability (CLAIM) compensation (CLAIM) severance pay (REPERCUSSION) overtime hours (CLAIM)
ADMISSIBILITY. APPEAL INTEREST. An ordinary appeal is not accepted when the appealing party has not been defeated concerning the chapter of the judgment subject to appeal.	APPEAL INTEREST (CLAIM) ordinary appeal (PROCEEDING_TYPE) not accepted (DECISION)	ordinary appeal (PROCEEDING_TYPE)

Table 11: Examples of sentences containing NER errors. The first sentence contains an example of boundary error for the "period of stability" entity, for which the model missed the two initial tokens. The model also predicted the "severance pay" entity as a repercussion instead of a claim. The entities "overtime due", "compensation", and "overtime hours" were correctly identified. The second sentence contains an example of a missed claim ("APPEAL INTEREST") and a missed decision ("not accepted"), as well as a correctly classified proceeding type ("ordinary appeal").

Sentences	Gold Relations	Predicted Relations
I acknowledge the ordinary appeal filed by the defendant and, on the merits, I grant it to reverse the judgment and dismiss the request for overtime and payment for the partially utilized intra-day interval.	(overtime, DECISION, dismiss) (intra-day interval, DECISION, dismiss)	(overtime, DECISION, reverse the judgment) (overtime, DECISION, dismiss)

Table 12: Examples of sentences containing RE errors. The sentence contains 6 labeled entities: four decisions and two claims. The gold and predicted relations presented in the table are in the format (subject entity, relation label, object entity). The model missed the relation between the "dismiss" decision and the "intra-day interval" claim, and predicted an unexisting relation between "reverse the judgment" and "overtime".