An Ensemble of LLMs finetuned with LoRA for NER in Portuguese legal documents

Rafael Oleques Nunes[1][0009−0007−8842−421X], Letícia Puttlitz[1][0009−0002−0313−1017], Antonio Oss Boll[1][0009−0004−5440−6126], Andre Spritzer[1][0009−0002−4232−1585], Carla Maria Dal Sasso Freitas[1][0000−0003−1986−8435], Dennis Giovani Balreira[1][0000−0002−0801−9393], and Anderson Rocha Tavares[1][0000−0002−8530−6468]

Federal University of Rio Grande do Sul, Brazil
{ronunes,spritzer,carla,dgbalreira,artavares}@inf.ufrgs.br
{leticia.puttlitz,antonio.boll}@ufrgs.br

Abstract. Given the high computational costs of traditional fine-tuning methods and the goal of improving performance,this study investigate the application of low-rank adaptation (LoRA) for fine-tuning BERT models to Portuguese Legal Named Entity Recognition (NER) and the integration of Large Language Models (LLMs) in an ensemble setup. Focusing on the underrepresented Portuguese language, we aim to examine the reliability of extractions enabled by LoRA models and glean actionable insights from the results of both LoRA and LLMs operating in ensembles. Achieving F1-scores of 88.49% for the LeNER-Br corpus and 81.00% for the UlyssesNER-Br corpus, LoRA models demonstrated competitive performance, approaching state-of-the-art standards. Our research demonstrates that incorporating class definitions and counting votes per class substantially improves LLM ensemble results. Overall, this contribution advances the frontiers of AI-powered legal text mining, proposing small models and initial prompt engineering to low-resource conditions that are scalable for broader representation.

Keywords: Named Entity Recognition · Large Language Models · LoRA.

1 Introduction

With the advances in Natural Language Processing (NLP), the domain of legal NLP also follows in constant evolution, facing similar challenges and obstacles. One of the main areas of work in the legal NLP field is Named Entity Recognition (NER), which aims to identify and extract entities in text, such as petitions, bills, and sentences from legal documents. The process of recognizing these entities has become challenging due to the lack of structure and predefined entities in corpora from the legal system. Another obstacle is to work in the Brazilian legal system domain, as it does not have the same number of documents and models as the English language.

To address these problems, researchers have created multiple corpora to tackle the issue of NER in the Brazilian legal domain, such as the LeNER-Br [5]

and UlyssesNER-Br [3], adding entities to multiple phrases and helping to create a legal NER-friendly corpus. Additionally, many BERT models were created to facilitate the Brazilian legal NER, including LegalBERT-pt [29], BERTikal [23], and many others.

In this work, we propose a comprehensive investigation centered on two primary objectives: first, to analyze the efficacy of deploying LoRA to fine-tune BERT architectures for NER tasks within the legal domain, in contrast with relevant literature; and second, to present a study on the design and implementation of prompt engineering tactics for LLM ensemble modeling, aiming to enhance robustness. Our core contributions are (i) delivering a study on the practicality and advantages of coupling LoRA and BERT models within the specialized context of legal NER, supplemented by comparison with related works, and (ii) an exposition on exploiting assorted prompt ingredients for building LLM ensembles.

2 Related Work

Natural language processing is continuously evolving with the introduction of new concepts. Each new technique opens doors to several fields of work, each with its unique modeling approach and output. Introduced by Mikolov in 2013 [18], the word embeddings gave the area of natural language processing (NLP) a new view, allowing words to be represented in a vectorized manner.

After that, in 2017, the Transformer model [31] revolutionized much of the work in the NLP area, inspiring the creation of several models based on its ideas, such as BERT [13] and GPT [24]. BERT, for instance, utilizes only the encoder part of the transformer and can understand context and work on several other tasks that require text comprehension. Meanwhile, GPT, utilizing the decoder, can generate text, among other tasks. Another model worth mentioning is the ELMo, introduced in 2018 [22], which represents terms in a vector space and makes them context-sensitive.

Researchers have been creating several domain-specific models in the last few years due to BERT's ability to pre-train on data. BERTimbau [30], for instance, was developed specifically for making inferences with Portuguese data, focusing mainly on Brazilian Portuguese corpora. It leverages the transformer architecture to achieve high perfomance in various natural language processing tasks such as Sentence Textual Similarity and Named Entity Recognition. For training, BERTimbau utilized the brWaC [32] corpus, which consists of a large Web corpus for Brazilian Portuguese. In the legal domain, several models have been developed for various languages, including English [10], German [12], Arabic [2], French [14], and Portuguese [29].

The corpora is also a crucial step in obtaining the results, as it needs to be from the domain and have the correct entities tied to it. A few of the legal corpora from the Brazilian legal system for Named Entity Recognition are LeNER-Br [5], UlyssesNER-Br [3], Brazilian Supreme Court corpus [11] and CDJUR-BR [9].

Concerning advances in models and the improvement of NER, previous studies have explored how transformer models can enhance NER performance. We focus on LeNER-Br and UlyssesNER-Br, as the former is predominantly a judiciary corpus, and the latter is a legislative corpus. This approach provides an interesting perspective on two different aspects of the legal domain.

Regarding the LeNER-Br corpus, Bonifácio et al. (2020) [6] found that using a domain-specific corpus during the pre-training of multilingual BERT can enhance the recognition of named entities. Additionally, Zanuz et al. (2022) [35] achieved state-of-the-art results by fine-tuning the LeNER-Br corpus using BERTimbau.

For the UlyssesNER-Br corpus, Albuquerque et al. (2023) [4] fine-tuned BERTimbau and conducted the first evaluation of this corpus using BERT models. Nunes et al. (2024) [20] investigated how a semi-supervised technique can improve BERTimbau's performance in the legislative domain, demonstrating that such techniques can enhance model results.

A recent study [21] focused on how a Generative Language Model (GLM) specialized in Brazilian Portuguese performs on both corpora. The authors found that In-Context Learning can provide initial results and demonstrated the potential of these models to extract entities. However, they noted that further studies are needed, as BERT models have shown better performance.

To the best of our knowledge, our work is the first to analyze the efficacy of using LoRA to fine-tune BERT models specifically for legal NER tasks and to explore the design and implementation of prompt engineering techniques for using LLM to ensemble legal NER models. These contributions offer a novel perspective on the study of legal NER and the practical advantages of combining LoRA and BERT models within this specialized domain.

3 Domain Corpora

This Section presents two types of corpora that recognize legally named entities, each focusing on a different type of data. They are called *LeNER-Br* [5] and *UlyssesNER-Br* [3], both sourced from the Brazilian legal system. We chose these two corpora because they can access different fronts of legal domains by leveraging judiciary and legislative texts. Both corpora are embedded within the legislative context. While sentences in LeNER-Br are extracted from court and legislative texts, those in UlyssesNER-Br come from legislative inquiries and bills from the Brazilian Chamber of Deputies.

The first corpus, introduced in 2018 and called *LeNER-Br* [5], focuses on data from the Brazilian justice system. It is composed of a total of 70 legal documents from various types of courts and legislations, containing a total of 10,392 sentences and 318,073 tokens. The categories of the named entities include "Person", "Legal cases", "Time", "Location", "Legislation" and "Organization".

Focusing on the legislative side of legal data, *UlyssesNER-Br* [3] consists of two types of data from the Brazilian Chamber of Deputies: legislative consultations (ST) and bills (PL). The ST database has 790 sentences and 77,441

tokens, whereas the PL corpus contains 9,526 sentences with 138,741 tokens. Both datasets include the following categories of named entities: “Fundamento", “Organizacao," “Produtodelei," “Local," “Data," and “Evento."

4 Models and fine-tuning with LoRA

This section provides an overview of the models used to recognize legally named entities. These models are variations of the BERT model [13], each pre-trained with a specific type of data, along with the addition of a Large Language Model.

BERTimbau [30], launched in 2020, is one of the first Brazilian BERT models. It was pre-trained using *brWaC* [33], a large corpus of data from the Brazilian web. Two types of BERTimbau were pre-trained: BERTimbau Base, with 12 layers, a hidden size of 768, 12 attention heads, and 110 million parameters; and BERTimbau Large, with 24 layers, a hidden size of 1024, 16 attention heads, and 330 million parameters. The maximum length for a sentence contains 512 tokens for both models. We used BERTimbau Base for the analysis. It was fine-tuned separately on the *UlyssesNER-Br* and *LeNER-Br* corpora. The hyperparameters used for this model are specified in Section 6.2.

Aiming to develop a legal-focused model, *LegalBERT-pt* [29] is a BERT model pre-trained on Brazilian legal corpora. The authors pre-trained two types of models: LegalBert-pt SC and LegalBert-pt FP. The SC model is formulated by pre-training a BERT model from scratch, with the same configuration as the BERTimbau base. Differently, the FP model pre-trains a BERTimbau-Base with the domain-specific corpora. The corpora used in the article consist of multiple Brazilian court cases and, for the SC model, Portuguese Wikipedia articles, totaling 1,500,000 legal documents and a vocabulary of 36,345 words. *LegalBERT-pt FP* was selected as the best-performing model.

In this study, *LegalBERT-pt FP* was fine-tuned using the corpora described in Section 3, with its hyperparameters detailed in Section 6.2.

Finishing the BERT variant models, *BERTikal* [23] is a BERT model trained on corpora from clippings, court cases and motions in the legal Brazilian data.

The previously described models are fine-tuned on LeNER-br and UlyssesNER-Br (see Section 3) and used hyperparameters described in Section 6.2. We perform our fine-tuning with LoRA [16]., which is a method to reduce the number of trainable parameters and hence the training computational cost using matrix algebra. The method freezes the pre-trained parameters of the model and optimizes the rank-decomposition of the dense layer matrices.

The only Large Language Model (LLM) we use is *Mistral 7B* [17]. It is composed of 7 Billion parameters, has a dimension size of 4096, 32 layers, a vocabulary size of 32000 and multiple other parameters and hyperparameters. We obtained the training corpus from the open web. As an LLM, the Mistral 7B model is capable of a diverse gamma of tasks, including knowledge, math, reasoning, comprehension, and code, among several others. We use this model for the ensemble technique explained in Section 5 and for a *Zero-Shot Learning* method.

5 An Ensemble of Language Models

Ensemble learning aims to improve the predictive performance of a single model by training multiple models and combining their predictions [26]. A simple form of ensemble is voting, where base classifiers are presented with an input and each makes a prediction. Subsequently, the prediction that receives the most votes is selected as the final output. We apply ensemble learning to combine multiple language models on the legal Named Entity Recognition task.

In our approach, BERT models serve as the base classifiers for Named Entity Recognition. After the BERT models generates their predictions, Mistral acts as an aggregator in the ensemble, combining the outputs. This method provides a unified result that incorporates and values all the predictions from the BERT models.

We based our prompt engineering on the following elements: persona, definitions, votes, answer format, sentence, and query. The prompt starts with a **persona**, which is used to impersonate a linguist specialized in political science to achieve the best possible performance [27]. We chose this role for the persona because it is a linguistic task that requires legal domain knowledge.

We used the **definitions** of the entity classes to alleviate the problem that some categories have non-intuitive meanings based on their names. For instance, the category *fundamento*[1] does not seem directly linked with legal norms or bills. Thus, we also give the entity class definitions in the prompt. Definitions were sourced from corpora [3,5] and translated into Portuguese. However, we adapted the definitions for *person* and *location* from Harem [28] because both corpora were influenced by Harem and adhered to its class definitions.

We designed the **votes** to provide information to the model regarding the classes given more or fewer votes from each classifier. We present the vote in the format *Class: x votes*, where *class* is the name of the entity class, and x is the number of votes.

In the **answer format**, we encourage the model to use the chain of thought (CoT) [34] by explaining the answer before providing the entity class. We divided the answers into *explanation* and *answer* to facilitate post-processing to obtain the class.

Finally, at the end of the prompt, we provide the **sentence** and ask the model to assign the right entity class to a term or say that it does not have a class in **question-answer** format. We provided entities from the classifiers in question, definitions, and votes. We also tested using all the entities for the definitions and questions.

All classes used in the prompt are given with the first letter in uppercase and the rest in lowercase. We also use space to divide n-gram terms, as in *produto de lei*. We opt to use this format so that the names remain in a more fluid language, providing a better context and avoiding unnecessary splitting in more tokens of the terms.

[1] In English: foundation.

Sometimes, the model gives answer classes near the goals but with slightly different names (e.g. typos). We used Sentence-BERT (SBERT) [25] for **post-processing**, along with the classes [21]. After comparison, we converted the answer into a vector annotating the entity of the term in the sentence. The vector is incrementally annotated because we use a prompt for each term.

6 Experimental Evaluation

This section provides an overview of the experimental environment, encompassing the principal libraries utilized, hyperparameters, and a thorough explanation of the hyperparameter tuning carried out for the LoRA models. In addition, we discuss the performance metrics utilized to evaluate our models.

6.1 Setup

We conducted the experiments on a computer with 12 GB RAM and an Nvidia GeForce RTX 4070 GPU. Owing to its broad support for libraries related to machine learning and natural language processing, we decided to implement our experiments using Python 3.7.6. Quantization, performed using the bitsandbytes library[2], was employed to reduce memory requirements and computational costs. The BERT and Mistral models were obtained from the HuggingFace Hub. The functions and methods of LoRA and running the models were used from HuggingFace.

6.2 Hyperparameters

In this section, we describe the experiment hyperparameters. We first present the hyperparameter search and final values for LoRA training. Here, we present the values used to train the BERT models. Finally, we present the hyperparameters for quantization and generate responses to the LLM.

LoRA. We used Optuna [1] to perform the hyperparameter search. We set the optimization range to drop out around 0.1 and 0.5, and the range of r and *alpha* between 16, 32, 64, and 128, with 100 iterations. The final hyperparameters were $r = 32$ *alpha* = 1.747406, *dropout* = 0.1 (without optimization) and bias = "all".

BERT. We followed previous studies that used the same legal corpora to set the hyperparameters for the BERT classifier [20,35,7]. The set of values was learning_rate = 1e-3, *batch_size* = 10, and *weight_decay_size* = 0.01.

LLM. We use 4-bit quantization to load the model in our machine, setting the attributes *load_in_4bit* = True, *bnb_4bit_compute_dtype* = torch.bfloat16, *torch_dtype* = torch.float16, *device_map* = "auto". To generate the answer, we used the standard value of the method *generate*, setting *max_new_tokens* = 800 and pad_token_id = *model.config.eos_token_id*.

[2] https://github.com/TimDettmers/bitsandbytes

6.3 Metrics

We used Seqeval [19] to compute the metrics. This library assesses the results by considering the sequence of tags assigned to each entity in a complete sentence, clearly recognizing entities beyond individual tokens. Our primary metric is the F1-Score, calculated for each class and overall. Additionally, precision and recall were computed for each class, whereas accuracy was determined for the overall results.

7 Results and Discussion

In this section, we present the outcomes of our experiments involving LoRa applied to BERT models and LLM as an ensemble approach. Initially, we delve into the performance of individual LoRa classifiers, considering any potential connections to the pretraining domain of the underlying base models. Subsequently, we reveal the combined results achieved through the LLM ensemble technique and scrutinize the relationships between the selected models and prompts. Finally, we compare our results with those existing state-of-the-art (SOTA) baselines.

7.1 LoRA Classifiers

Table 1. F1-Score for each corpus to the LoRA classifiers.

Model	LeNER-Br	UlyssesNER-Br
LegalBERT-pt	**88.49**	76.96
BERTimbau	87.12	**81.00**
BERTikal	81.53	67.72

We present the final F1-Score of the classifiers to each corpus in Table 1. LegalBERT-pt presented the best result for LeNER-BR, obtaining 88%, followed by BERTimbau, with 87%. For UlyssesNER-Br, we obtained the inverse result, where BERTimbau had the best result at 81%, followed by LegalBERT-pt with 76%. This result probably occurred because LegalBERT-pt follows BERTimbau pretraining in judiciary documents (more information in Section 4), which is the data domain of LeNER-Br, whereas UlyssesNER-Br uses legislative documents, which can explain the lower results in this case.

However, it is important to note that we cannot express the consistency of the difference between the close results because we did not use k-fold cross-validation to obtain the average result and variance around the folds. Furthermore, we did not compute tests to obtain statistical significance.

Another observation regards the response format, which we discovered during manual inspection. Specifically, instances exist where the generated outputs fail to adhere strictly to the prescribed formats. Given this discrepancy, there is a risk that accurate responses might go undetected due to unexpected presentation, consequently contributing to reduced F1 scores.

7.2 LLMs Applied as Ensemble

Table 2. Overall F1-Score for UlyssesNER-Br using Mistral 7B as ensembler.

Model	Prompt	F1-Score
BERTimbau	definitions + votes	69.87
BERTimbau	definitions	67.97
legalbert	definitions + votes	64.30
BERTimbau	definitions + votes + all labels	63.81
legalbert	definitions	62.15
BERTimbau + legalbert	definitions	60.82
BERTimbau + legalbert	definitions + votes	59.32
legalbert	definitions + votes + all labels	58.36
legalbert + bertikal	definitions	57.83
BERTimbau + bertikal	definitions + votes	57.60
BERTimbau + bertikal	definitions	57.50
BERTimbau + legalbert	definitions + votes + all labels	56.05
legalbert + bertikal	definitions + votes	55.21
BERTimbau + legalbert + bertikal	definitions + votes	55.16
BERTimbau + legalbert + bertikal	definitions	53.51
BERTimbau + bertikal	definitions + votes + all labels	52.70
BERTimbau	definitions + all labels	51.95
legalbert + bertikal	definitions + votes + all labels	50.20
BERTimbau + legalbert + bertikal	definitions + votes + all labels	47.80
legalbert	definitions + all labels	47.09
bertikal	definitions + votes	46.76
BERTimbau + legalbert	definitions + all labels	45.99
bertikal	definitions	45.95
bertikal	definitions + votes + all labels	44.97
BERTimbau + legalbert + bertikal	definitions + all labels	43.65
BERTimbau + bertikal	definitions + all labels	42.28
legalbert + bertikal	definitions + all labels	40.48
bertikal	definitions + all labels	34.78

The performance resulting from employing Mistral 7B as an ensemble approach can be found in Tables 2 and 3. We conducted experiments covering diverse combinations of prompts and models to evaluate the effect of each model-prompt pairing. Surprisingly, the combination of multiple models did not yield improved outcomes. The top three scores across both datasets featured BERTimbau and LegalBert-Pt employed independently, echoing the trend seen in Table 1, wherein BERTimbau outperformed UlyssesNER-Br, whereas LegalBert-Pt proved superior when working with LeNER-Br.

Regarding prompt engineering, our investigation revealed that utilizing definitional descriptions paired with voting yielded the most successful outcomes. This strategy offers the model clear class definitions and individual class preferences as indicated by the participating models. Although the label distribution

Table 3. Overall F1-Score for LeNER-Br using Mistral 7B as ensembler.

Model	**Prompt**	**F1-Score**
legalbert	definitions + votes	63.03
BERTimbau	definitions + votes	62.38
legalbert	definitions + votes + all labels	61.71
legalbert + bertikal	definitions + votes	61.18
BERTimbau	definitions + votes + all labels	61.07
legalbert	definitions	60.99
BERTimbau + bertikal	definitions + votes	60.74
BERTimbau	definitions	60.60
BERTimbau + legalbert	definitions + votes	60.46
BERTimbau + legalbert	definitions + votes + all labels	60.13
legalbert + bertikal	definitions + votes + all labels	60.12
BERTimbau + legalbert + bertikal	definitions + votes	59.57
legalbert + bertikal	definitions	59.31
BERTimbau + bertikal	definitions + votes + all labels	59.22
BERTimbau + bertikal	definitions	59.11
BERTimbau + legalbert	definitions	58.94
BERTimbau + legalbert + bertikal	definitions + votes + all labels	57.62
BERTimbau + legalbert + bertikal	definitions	57.41
legalbert	definitions + all labels	51.86
BERTimbau	definitions + all labels	50.55
legalbert + bertikal	definitions + all labels	50.24
BERTimbau + legalbert	definitions + all labels	49.86
BERTimbau + bertikal	definitions + all labels	49.33
BERTimbau + legalbert + bertikal	definitions + all labels	48.59
bertikal	definitions + votes	5.26
bertikal	definitions + votes + all labels	5.18
bertikal	definitions	4.91
bertikal	definitions + all labels	3.65

appears scattered throughout the findings, the majority converge near 50% for LeNER-Br and fall below this threshold for UlyssesNER-Br. These observations underscore the significance of defining a constrained solution space when using LLMs, facilitating enhanced categorization capabilities.

Considering the exploratory findings presented thus far, it is important to highlight the necessity of performing more deliberate examinations to acquire robust statistical comparisons between various prompt structures and models, as discussed in Section 7.1. Furthermore, embracing a multifaceted validation strategy, including repetitive experimentation (K iterations), calculation of mean values, and determination of standard deviations, constitutes another key element in fostering greater certainty surrounding the obtained outcomes.

7.3 Our Results and the State-of-the-art

Table 4. Overall F1-Score for each corpus for our classifiers and SOTA classifiers. The results for UlyssesNER-Br are from Nunes et al. (2024) [20], and the results for LeNER-Br are from Zanuz et al. (2022) [35].

Model	**UlyssesNER-Br**	**LeNER-Br**
BERTimbau + self-learning	**86.70 ± 2.28**	-
BERTimbau	83.53 ± 2.56	**91.14 ± 0.39**
LoRA (ours)	81.00	88.49
LLM (ours)	69.86	63.03

Table 5. F1-Score for each class in UlyssesNER-Br for our classifier and from Nunes et al. (2024) [20]. The underlined values are those near the established results.

Category	**BERTimbau + LoRA**	**BERTimbau + Self-Learning**
PRODUTODELEI	57.14	75.42 ± 4.47
PESSOA	<u>89.60</u>	87.48 ± 2.79
ORGANIZACAO	74.01	84.89 ± 5.77
LOCAL	74.37	86.46 ± 3.73
FUNDAMENTO	<u>85.06</u>	88.60 ± 2.29
EVENTO	<u>47.06</u>	58.10 ± 34.16
DATA	<u>97.49</u>	94.77 ± 2.65

Comparing our results with those of the SOTA, Table 4 reveals that the LoRA classifiers achieved similar performance to previous studies [20,35]. Specifically, our LoRA classifier obtained results closely aligned with a classifier solely utilizing BERTimbau [20] on the UlyssesNER-Br corpus and approached the performance of the classifier proposed by Zanunz et al. (2022) [35] on the LeNER-Br corpus.

Table 6. F1-Score for each class in LeNER-Br for our classifier and from Zanuz et al. (2022) [35]. The underlined values are those near the established results.

Category	**LegalBert-Pt + LoRA**	**BERTimbau**
TEMPO	92.02	96.04 ± 0.58
PESSOA	<u>95.86</u>	97.38 ± 0.44
ORGANIZACAO	<u>86.04</u>	86.66 ± 1.17
LOCAL	<u>73.58</u>	75.67 ± 3.18
LEGISLACAO	92.66	95.90 ± 0.83
JURISPRUDENCIA	78.66	87.76 ± 0.87

The classes that achieved results similar to those of Nunes et al. (2024) [20] were *PESSOA*, *FUNDAMENTO*, *EVENTO*, and *DATA*, as shown in Table 5. Notably, the structuring of entities, such as *DATA* (date) and *PESSOA* (person), presents a plausible explanation for their relatively higher performance. Dates typically adhere to specific formats, facilitating their distinction from other entities, whereas the names of individuals often follow recognizable patterns similar to those observed in *FUNDAMENTO* (laws). Conversely, *EVENTO* (event) posed a more significant challenge because of the limited training and test examples available for this class.

Furthermore, a comparative analysis with Zanuz et al. (2022) [35] revealed that classes such as *PESSOA*, *ORGANIZACAO*, and *LOCAL* demonstrate similar performance trends, as shown in Table 6. This alignment could be attributed to factors such as the extensive training, validation, and test data available for each class and the inherent structural complexities associated with entities in these categories, as discussed previously.

These results show the power of LoRA for the NER task, with the capacity to achieve good results by training a smaller number of parameters, i.e., we decrease the number of training params from 109,532,186 to only 1,291,789 params. Thus, this approach results in small and faster models that can be used with less computer power and stored in less space.

8 Conclusion

In this work, we presented a study of using LoRA to fine-tune BERT Portuguese models to NER tasks in the legal domain. The LoRA models could achieve near results compared to state-of-art for each corpora tested. Another point is that we have the advantage of training smaller models, which is an advantage to storing and running the model in machines with low processing.

We also tested using Mistral 7B applied to the ensemble, which was unproductive. The results were lower compared to the LoRA models. However, we recognize that some outputs do not respect the format requested, which can be a point that the right answers are not recognized.

In future work, we aspire to incorporate hyperparameter tuning tailored to each model augmented with LoRa, potentially amplifying the overall perfor-

mance. Moreover, we intend to utilize k-fold cross-validation to derive more authoritative and comparable results than existing literature.

Regarding ensembling, we envision expanding our scope by exploring alternative prompts engineered using prompt manipulation techniques. Complementarily, we seek to develop post-processing routines accommodating minor fluctuations in the output format, thereby minimizing false negatives and boosting performance indicators. Testing other multilingual LLMs (e.g., LLaMA2) and trialing retrieval mechanisms, such as In-Context Learning, are other possible ways to improve results.

Another interesting future approach is to compare our results with classic ensemble strategies, such as bagging [8] and boosting [15]. Simple fusion schemes such as averaging and maximum selection may provide promising avenues warranting investigation in this dynamic domain.

Acknowledgements. This work has been partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also acknowledge financial support from the Brazilian funding agency CNPq.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. AL-Qurishi, M., AlQaseemi, S., Soussi, R.: Aralegal-bert: A pretrained language model for arabic legal text (2022)
3. Albuquerque, H.O., Costa, R., Silvestre, G., Souza, E., da Silva, N.F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., et al.: Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In: International Conference on Computational Processing of the Portuguese Language. pp. 3–14. Springer (2022)
4. Albuquerque, H.O., Souza, E., Oliveira, A.L., Macêdo, D., Zanchettin, C., Vitório, D., da Silva, N.F., de Carvalho, A.C.: On the assessment of deep learning models for named entity recognition of brazilian legal documents. In: EPIA Conference on Artificial Intelligence. pp. 93–104. Springer (2023)
5. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R., Stauffer, M., Couto, S., Bermejo, P.: LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In: Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13. pp. 313–323. Springer (2018)
6. Bonifacio, L.H., Vilela, P.A., Lobato, G.R., Fernandes, E.R.: A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9. pp. 648–662. Springer (2020)

7. Bonifacio, L.H., Vilela, P.A., Lobato, G.R., Fernandes, E.R.: A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9. pp. 648–662. Springer (2020)
8. Breiman, L.: Bagging Predictors. Machine Learning **24**, 123–140 (1996). https://doi.org/10.1007/BF00058655
9. Brito, M., Pinheiro, V., Furtado, V., Neto, J.A.M., Bomfim, F.d.C.J., da Costa, A.C.F., Silveira, R.: Cdjur-br-uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In: Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. pp. 177–186. SBC (2023)
10. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-bert: The muppets straight out of law school (2020)
11. Correia, F.A., Almeida, A.A., Nunes, J.L., Santos, K.G., Hartmann, I.A., Silva, F.A., Lopes, H.: Fine-grained legal entity annotation: A case study on the brazilian supreme court. Information Processing & Management **59**(1), 102794 (2022)
12. Darji, H., Mitrović, J., Granitzer, M.: German bert model for legal named entity recognition. In: Proceedings of the 15th International Conference on Agents and Artificial Intelligence. SCITEPRESS - Science and Technology Publications (2023). https://doi.org/10.5220/0011749400003393, http://dx.doi.org/10.5220/0011749400003393
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)
14. Douka, S., Abdine, H., Vazirgiannis, M., Hamdani, R.E., Amariles, D.R.: Juribert: A masked-language model adaptation for french legal text (2022)
15. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. Machine learning: Proceedings of the thirteenth international conference **96**, 148–156 (1996)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
17. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B (2023)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
19. Nakayama, H.: seqeval: A python framework for sequence labeling evaluation (2018), https://github.com/chakki-works/seqeval, software available from https://github.com/chakki-works/seqeval
20. Nunes, R.O., Balreira, D.G., Spritzer, A.S., Freitas, C.M.D.S.: A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese. pp. 290–300 (2024)
21. Oleques Nunes., R., Spritzer., A., Dal Sasso Freitas., C., Balreira., D.: Out of sesame street: A study of portuguese legal named entity recognition through in-context learning. In: Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS. pp. 477–489. INSTICC, SciTePress (2024). https://doi.org/10.5220/0012624700003690
22. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations (2018)

23. Polo, F.M., Mendonça, G.C.F., Parreira, K.C.J., Gianvechio, L., Cordeiro, P., Ferreira, J.B., de Lima, L.M.P., do Amaral Maia, A.C., Vicente, R.: LegalNLP-Natural Language Processing methods for the Brazilian Legal Language. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. pp. 763–774. SBC (2021)
24. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. OpenAI Technical Report (2018)
25. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2020), https://arxiv.org/abs/2004.09813
26. Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley interdisciplinary reviews: data mining and knowledge discovery **8**(4), e1249 (2018)
27. Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., Akata, Z.: In-context impersonation reveals large language models' strengths and biases. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=CbsJ53LdKc
28. Santos, D., Cardoso, N.: A golden resource for named entity recognition in portuguese. In: International workshop on computational processing of the portuguese language. pp. 69–79. Springer (2006)
29. Silva, N., Silva, M., Pereira, F., Tarrega, J., Beinotti, J., Fonseca, M., Andrade, F., Carvalho, A.: Evaluating Topic Models in Portuguese Political Comments About Bills from Brazil's Chamber of Deputies. In: Anais da X Brazilian Conference on Intelligent Systems. SBC, Porto Alegre, RS, Brasil (2021), https://sol.sbc.org.br/index.php/bracis/article/view/19061
30. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT Models for Brazilian Portuguese, pp. 403–417 (10 2020). https://doi.org/10.1007/978-3-030-61377-8_28
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need (2017)
32. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brwac corpus: A new open resource for brazilian portuguese. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018) (2018)
33. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brWaC corpus: A new open resource for Brazilian Portuguese. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1686
34. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
35. Zanuz, L., Rigo, S.J.: Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 219–229. Springer (2022)