Out of Sesame Street: A Study of Portuguese Legal Named Entity Recognition Through In-Context Learning

Rafael Oleques Nunes[a], Andre Suslik Spritzer[b], Carla Maria Dal Sasso Freitas[c]
and Dennis Giovani Balreira[d]
Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Keywords: In-Context Learning, Large Language Models, Named Entity Recognition, Legal Tech, LLama.

Abstract: This paper explores the application of the In-Context Learning (ICL) paradigm for Named Entity Recognition (NER) within the Portuguese language legal domain. Identifying named entities in legal documents is complex due to the intricate nature of legal language and the specificity of legal terms. This task is important for a range of applications, from legal information retrieval to automated summarization and analysis. However, the manual annotation of these entities is costly due to the specialized knowledge required from legal experts and the large volume of documents. Recent advancements in Large Language Models (LLM) have led to studies exploring the use of ICL to improve the performance of Generative Language Models (GLMs). In this work, we used Sabiá, a Portuguese language LLM, to extract named entities within the legal domain. Our goal was to evaluate the consistency of these extractions and derive insights from the results. Our methodology involved using a legal-domain NER corpus as input and selecting specific samples for a prompting task. We then instructed the GLM to catalog its own NER corpus, which we compared with the original test examples. Our study examined various aspects, including context examples, selection strategies, heuristic methodologies, post-processing techniques, and quantitative and qualitative analyses across specific domain classes. Our results indicate promising directions for future research and applications in specialized domains.

1 INTRODUCTION

The inherent complexity of legal language, coupled with the specificity of its terminology, makes recognizing named entities in legal documents an interesting challenge within the field of natural language processing. The significance of this task reverberates across various applications, from facilitating legal information retrieval to enabling automated summarization and in-depth analysis, unlocking opportunities for further exploration and understanding.

Despite its undeniable benefits, the manual annotation of named entities in legal documents is a resource-intensive process. The specialized knowledge required from legal experts, combined with the sheer volume of legal documents, makes conventional annotation approaches costly and time-consuming. To address these issues, the integration of Large Language Models (LLM) has emerged as a transformative force, offering the promise of automating the process and enhancing the accuracy of named entity recognition (Barale et al., 2023).

Within this context, In-Context Learning (ICL) stands out as a particularly promising paradigm, offering several advantages (Dong et al., 2022). These include: (i) the simplicity of incorporating knowledge through a natural language interface (Brown et al., 2020), (ii) the capacity of learning from analogy (Winston, 1980), and (iii) the adeptness in handling large-scale real-world tasks (Sun et al., 2022). Given the increasing and successful adoption of In-Context Learning for several tasks, we set out to investigate how the application of this method to GLM can improve the NER task for Portuguese language legal corpora.

In this work, we used Sabiá, a Large Language Model (LLM) for the Portuguese language, to extract named entities within the legal domain. Our main goal was to evaluate the consistency of these extractions and derive insights from the results obtained. Additionally, we investigated three heuristics to re-

[a] https://orcid.org/0009-0007-8842-421X
[b] https://orcid.org/0009-0002-4232-1585
[c] https://orcid.org/0000-0003-1986-8435
[d] https://orcid.org/0000-0002-0801-9393

Nunes, R., Spritzer, A., Freitas, C. and Balreira, D.
Out of Sesame Street: A Study of Portuguese Legal Named Entity Recognition Through In-Context Learning.
DOI: 10.5220/0012624700003690
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 1*, pages 477-489
ISBN: 978-989-758-692-7; ISSN: 2184-4992
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

trieve in-context examples with the goal of improving ICL results. Finally, we explored different methods to mitigate the noise produced by misspellings and out-of-domain classes predicted by the LLM.

The remaining sections are organized as follows. Section 2 delves into the related work. Section 3 details our methodology, describing the legal corpora, the Portuguese language LLM and the prompt template that were used, the heuristics that were examined, and the post-processing filters that were applied. Section 4 describes the hardware setup employed in our experiments, the hyperparameters used, the division of the corpora, and the metrics used in the evaluation. Section 5 presents both quantitative and qualitative analyses of the results, accompanied by an investigation of the effect of postprocessing filters. Finally, Section 6 concludes the study with a discussion of the final results and perspectives for future research.

2 RELATED WORK

We summarize the most relevant studies divided into three topics: (i) methods of In-Context Learning; (ii) Generative LLMs for the Portuguese language; and (iii) works specific to the legal domain.

2.1 In-Context Learning

LLMs' growing prevalence and capability have led to many new studies in NLP that take the In-Context Learning paradigm (Dong et al., 2022). These works investigate several key points in this domain, such as prompt retrieval (Rubin et al., 2021), example selection (Zhang et al., 2022; Ye et al., 2023; Liu et al., 2021), informative extraction (Li and Qiu, 2023; Gupta et al., 2023), and even label bias mitigation (Fei et al., 2023).

Despite addressing NER-related tasks, such as entity locating and typing (Shen et al., 2023; Barale et al., 2023) and showing promising results, the above models are not explicitly focused on NER.

Other studies are focused specifically on NER. Among these, Ziyadi et al. (2020), Cui et al. (2021), and Huang et al. (2020) propose few-shot learning approaches. Other works include attempts to improve entity typing pipelines, with specific attention to broadening the scope of entities beyond conventional categories (Choi et al., 2018; Dai et al., 2021), and handling entities not encountered during training (Epure and Hennequin, 2022; Lin et al., 2020). Concerning corpora in the legal domain, Barale et al. (Barale et al., 2023) explore the retrieval of knowledge by evaluating entity typing as a proxy for assessing legal comprehension, comparing various legal-specific and generic LMs and prompting methods. Findings reveal that while LLama2 shows potential with optimized templates, law-oriented LMs exhibit inconsistent performance, and all models face challenges with entities from specific legal sub-domains.

To the best of our knowledge, we are the first to investigate this particular field in the Portuguese language. Our domain follows Barale's work (Barale et al., 2023) as well as the significant findings using the In-Context Learning approach in other contexts (Ye et al., 2023; Gupta et al., 2023; Feng et al., 2023).

2.2 Generative LLMs for Text

Generative Language Models (GLMs) are a class of artificial intelligence models designed to generate human-like text based on the input they receive. They use natural language processing and machine learning techniques to understand and produce coherent and contextually relevant sentences. Recently, they have gained special attention due to their high performance on contextual understanding, high flexibility, and generalization (Ahuja et al., 2023), mainly because of the transformer architecture (Vaswani et al., 2017). Despite these models using most prominently English corpora, several are multilingual, i.e., supporting other languages. Some of the most prominent multilingual models (and variations thereof) include Generative Pre-trained Transformer (GPT) (Brown et al., 2020), Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al., 2019), Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), Large Language Model Meta AI (LLama) (Touvron et al., 2023), and Mistral (Jiang et al., 2023). Refer to Amatriain et al. (Amatriain et al., 2023) for a comprehensive yet simple catalog of transformer models for further examples.

While all these models support multiple languages, only Sabiá (Pires et al., 2023a) incorporates specific corpora in Portuguese. This model, a large language model pre-trained on Portuguese, exhibits a slightly better performance than ChatGPT-3.5 across 14 Portuguese datasets. To the best of our knowledge, as a model that is inherently fine-tuned for Portuguese, it is the *only* model that is fine-tuned for this language.

2.3 Legal Domain NER Corpora

As of this writing, there are two available datasets that can be used for Portuguese language legal domain NER.

The LeNER-Br (Luz de Araujo et al., 2018) dataset comprises legal documents and includes tags for persons, locations, time entities, organizations, law, and legal cases. Its authors have demonstrated its effectiveness with retrained LSTM-CRF models, achieving good overall results for legislation and legal case entities.

The second one, UlyssesNER-Br (Albuquerque et al., 2022), contains two sources of information divided into two corpora for each reference source, including bills and legislative consultations. The proposed corpus categorizes entities into two types: *category* and *type*. Categories include conventional entities as well as additional legislative entities, while types represent specific instances within categories in accordance with a hierarchy. Since our work delves into the legal domain, we investigate both datasets.

3 METHODOLOGY

In this section, we delineate the proposed methodology across several key subsections following the pipeline shown in Figure 1. Subsection 3.1 describes the essential corpora used in this study. Subsection 3.2 provides a succinct description of the selected LLM. Subsection 3.3 explains the template employed and its composition. Subsection 3.4 intricately explores crafted heuristics, particularly emphasizing retrieval in In-Context Learning. Finally, Subsection 3.5 elaborates on the filters that play a crucial role in refining the NER model's outputs within the legislative context. This methodology is designed to provide comprehensive guidelines for evaluating NER performance in legal documents using In-Context Learning techniques.

3.1 Corpora

For a comprehensive understanding of In-Context Learning within the legal domain, we used two distinct NER corpora: LeNER-Br (Albuquerque et al., 2022) and UlyssesNER-Br (Luz de Araujo et al., 2018), which we discuss next. The former is dedicated to the judiciary domain, comprising Brazilian court documents, while the latter focuses on the legislative domain, comprising legislative bills.

LeNER-Br: (Luz de Araujo et al., 2018) is the first manually annotated Brazilian Legal corpus. It consists of 70 documents, of which 66 are from courts and tribunals and four from legislative documents. Table 1 lists the entity types and the number of entities for each type in each training, validation, and testing set.

Table 1: Frequency of named entities in LeNER-Br for each entity type.

Entity type	**Train**	**Valid**	**Test**
PESSOA	4,612	894	735
JURISPRUDENCIA	3,967	743	660
TEMPO	2,343	543	260
LOCAL	1,417	244	132
LEGISLACAO	13,039	2,609	2,669
ORGANIZACAO	6,671	1,608	1,367
Total	31,049	6,641	5,823

Table 2: Frequency of named entities in UlyssesNER-Br for each entity type.

Entity type	**Train**	**Valid**	**Test**
DATA	433	72	98
PESSOA	628	114	119
ORGANIZACAO	435	81	94
FUNDAMENTO	490	107	124
LOCAL	369	145	101
PRODUTODELEI	230	46	54
EVENTO	9	5	9
Total	2,594	570	599

UlyssesNER-Br: (Albuquerque et al., 2022) is a Brazilian Legislative corpus that includes 9,526 manually annotated sentences from 150 bills (*projetos de lei*) of the Brazilian Chamber of Deputies (the lower house of Brazil's National Congress). This corpus has two levels of entity types: *category* and *type*. As pointed out by the authors, the type level does not provide much additional information for model learning. As such, to allow for a better comparison with LeNER-Br entities, we opt to use only the category level. Entity categories and the number of examples each respectively contains are listed in Table 2.

3.2 Model

In this study, we used Sabiá (Pires et al., 2023b), a Portuguese language fine-tuned LLM trained on the ClueWeb 2022 dataset (Overwijk et al., 2022a,b). The training process involved expanding the capabilities of LLama 7B, LLama 65B, and GPT-J models. Our implementation specifically leveraged Sabiá-7B, which is based on the Sabiá model fine-tuned using LLama 7B.

3.3 Prompt Template

The template used in ICL holds significant importance as it efficiently embeds human knowledge into LLMs without requiring a fine-tuning process involving the incorporation of related examples within a

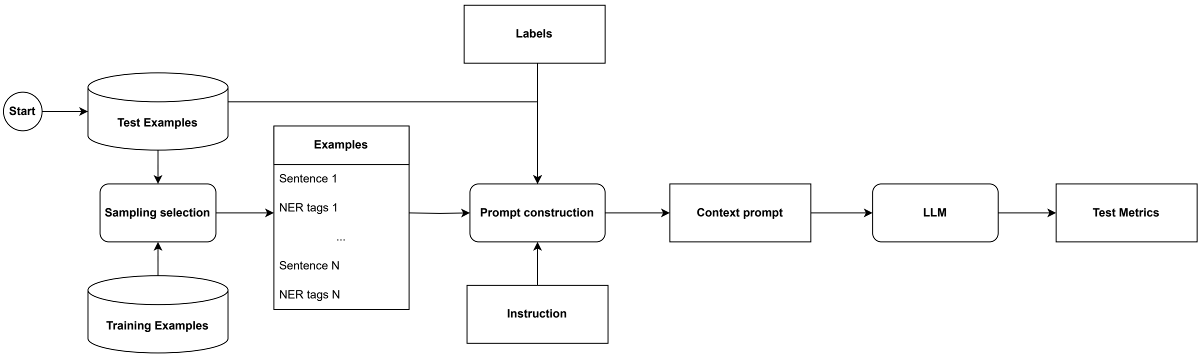


Figure 1: Pipeline showing our In-Context Learning NER approach. Given a selection of test examples, followed by the extraction of samples from the training set for ICL, we build a prompt using predefined instructions and target labels. We then feed the prompt into the LLM, where inference occurs. Finally, we derive the computed metrics from the output of the model.

prompt or altering its templates (Dong et al., 2022). We segmented our template into four key components to accomplish this objective and leverage existing knowledge: command, list of entities, examples, and input.

The template structure, comprising direct commands, examples, and input goals draws inspiration from previous studies (Kew et al., 2023; Barale et al., 2023; Fei et al., 2023; Feng et al., 2023). The inclusion of goal-oriented entities was essential for the experiments. Many inferences mislabel entities, categorizing them as types outside the scope or as parts of speech. This underscores a critical issue in NER, particularly with diverse entity classes, as it indicates that in-context examples alone may not be enough. Consequently, our template offers a comprehensive natural language description, such as *As categorias possíveis são:* {*categories*} (i.e., "the possible categories are: {categories}"), encompassing all possible entity types (Barale et al., 2023) and thus making it easier for each entity to refer to one of the desired types.

We structured our examples and input in {Sentence-Terms} format. This format pairs sentences with their respective terms and categories, which are represented as *Termos com categorias* (i.e., "terms with categories") to emphasize to the LLM that the expected output includes the term along with its associated category (as expressed in the category list). The terms and their respective categories were formatted as {*term*} *eh um* {*category*} (i.e., "{term} is a {category}"). In the initial tests, we experimented with mapping terms and their categories to as *(term, category)* tuples, but answers given by the model were of lower quality, leading us to adopt the current template format, which provided satisfactory results.

We opted to use the command *Reconheça os termos significativos e suas categorias* ("Identify significant terms and their categories"), prioritizing the term "categories" instead of using "named entities". Initial testing revealed the latter led to subpar results due to increased noise. The former command proved the most effective template for retrieving better and more accurate results.

3.4 In-Context Learning Heuristics

In the selection of in-context examples, we employed three distinct heuristics: (i) the top K similar examples, (ii) the top K similar examples chosen by entity type, and (iii) K random samples. Additionally, across all these heuristics, we conducted experiments to examine the significance of the example order, testing both ascending and descending orders based on the similarity with the input sentence.

***K* Similars.** Following prior research (Liu et al., 2021), we adopted a strategy of selecting the K most similar examples from the training set to act as contextual references, aiding the LLM in accurately associating entities with their respective terms. To accomplish this, we used Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) and the "distiluse-base-multilingual-cased" multilingual model (Reimers and Gurevych, 2020) to generate sentence embeddings using a semantic search. The cosine distance metric was employed to gauge the similarity between the embeddings.

***K* Similars per Entitiy Type.** Similar to the *K similars* heuristic, we retrieved the K most similar examples to the input. In this case, however, we ensured the inclusion of at least one example from each entity type. In scenarios where K exceeds the number of types, additional examples are selected based on their similarity without considering type affiliation. The primary objective of this heuristic is to investigate whether incorporating greater diversity across entity types can enhance classification results. By ensuring representation from various types, we aim to verify whether the model's classification performance benefits from a more comprehensive exposure to diverse instances.

Random Sampling. We employed random sampling to assess the significance of contextual examples in information extraction, enabling a comparison of the importance of in-context examples that only have goal-specific entity types with those that demonstrate higher similarity. By evaluating these factors, we aimed to determine whether the presence of random examples from goal-specific entity types was more important than the level of similarity for successful information extraction.

3.5 Filtering

Despite providing templates for the LLM containing the goal types, unexpected values could still be observed in the results. To address this, we implemented a series of filters. Initially, we eliminated punctuation from candidate entity types. Next, we used multilingual SBERT (Reimers and Gurevych, 2020) to obtain the cosine similarity between candidate and true types. We then selected the most similar types by retaining those that were higher or equal to a predefined threshold and disregarding the others. Our method only used the resultant value if the type was deemed valid; otherwise, we classified the token as 'O', indicating an unspecified entity.

4 EXPERIMENTAL EVALUATION

In this section, we present a detailed account of the experimental evaluations conducted to assess the proposed approach. This includes a comprehensive description of the environmental setup, outlining the hardware configurations and the libraries used. Furthermore, we detail the specific model, heuristics, and filter hyperparameters employed in our evaluation. In addition, we delineate the segmentation of the corpora and explain the metrics used to evaluate the obtained results.

4.1 Setup

Our heuristics, inference, and filtering tasks were executed in a computer equipped with an Nvidia GeForce RTX 4070 GPU and 12 GB of RAM. We chose the Python 3.7.6 as our programming language because of its extensive support for Machine Learning and Natural Language Processing libraries.

For the NLP models, we used the Maritalk API[1] to access the Sabiá model and Sentence Transformers library[2] to perform the semantic search. The seqeval (Nakayama, 2018) library was used to compute the NER metrics.

4.2 Hyperparameters

Model Hyperparameters. We adopted the recommended parameters for In-Context Learning[3] outlined in the official documentation of Sabiá. We set the *max_tokens* parameter to 1,000 to allow extensive responses. This choice facilitated longer template answers, as described in subsection 3.3, enriched with additional information, enabling the retrieval of answers related to all the entities.

Heuristic Hyperparameters. We used a range of K values from 1 to 8 for the heuristics in the LeNER-Br corpus, with the exception of the *k similar per entity type* heuristic, for which the range was narrowed from 5 to 8. Similarly, for the UlyssesNER-Br corpus, we used a range of K values spanning from 1 to 14, and for *k similar per entity type*, we set the range between 4 and 10. The variation in the number of examples between the two corpora stemmed from the larger size of the LeNER-Br, requiring more extensive training and validation periods. Nonetheless, both corpora exhibited comparable results, and the chosen number of examples sufficed for inter-corpora analysis, as elaborated in Section 5. To ensure reproducibility, we set the random seed to 42 for both corpora using random sampling.

Filters Hyperparameters. Throughout the empirical evaluation, we applied a cosine similarity threshold of 0.95, which proved to be effective in filtering. Additionally, we limited punctuation management to the character '-', given that the input was structured in the BIO format.

4.3 Corpora Division

We adhered to the standard division provided by the authors of the corpora, which includes training, validation, and test sets for each corpus. Tables 1 and 2 list the divisions used.

[1] https://maritaca-ai.github.io/Maritalk-api/Maritalk.html

[2] https://www.sbert.net/

[3] https://maritaca-ai.github.io/Maritalk-api/Maritalk.html

The training set was leveraged to retrieve in-context examples, while the test set was used for metric evaluation.

4.4 Metrics

To evaluate overall performance as well as the performance of each entity class, we used the micro F1, precision, and recall metrics, which were computed with the seqeval (Nakayama, 2018) library. A notable feature of this library is its metric calculation method, which relies on a sequence of tags assigned to each entity. This approach ensures that the accurate recognition of an entity in a complete sentence is prioritized over the simple identification of individual tokens.

5 RESULTS AND DISCUSSION

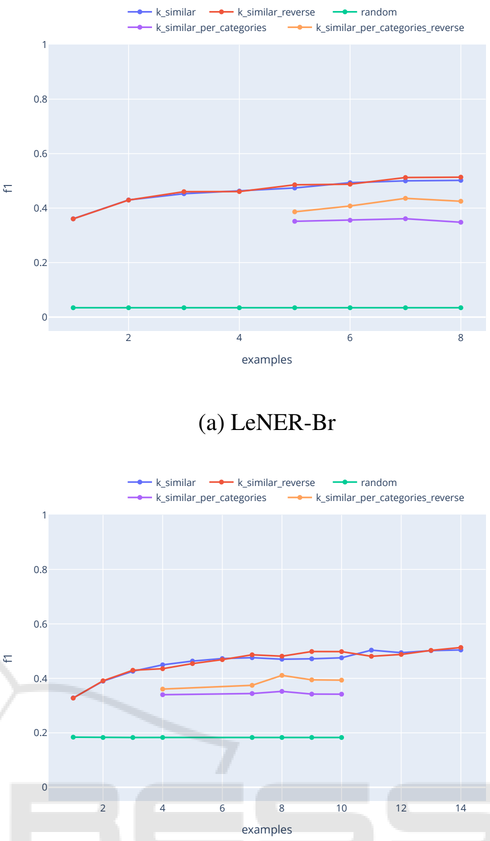
This section presents an in-depth exploration of the heuristics results, encompassing global and local perspectives as well as a comprehensive analysis of mispredictions associated with each entity and a description of the discernible impact of individual filters on overall and specific entity outcomes.

5.1 Heuristics Results

We present our heuristics results according to global and local perspectives for both LeNER-Br and UlyssesNER-Br. In global results, we analyze how each heuristic fared across *all* entities, whereas in local results we consider each entity separately.

An important point about the experiments is that we did not use the same number of iterations for all heuristics due to the time and cost involved. We followed the experiments in the UlyssesNER-Br corpus to discover when the same LeNER-Br F1-Score could be enhanced. Thus, we defined the same number to start the K similar examples per entity type for each corpus, as it would demonstrate similar results to values lower than the number of categories.

Global Results. The results shown in Figure 2 offer a summary comparison of each corpus, demonstrating the effect the number of examples in the prompt has on the F1-Score in ICL. Notably, the LeNER-Br corpus achieved 51% of F1-Score using eight examples, whereas UlyssesNER-Br required eleven examples to reach the same percentage. Moreover, UlyssesNER-Br attained 48% of F1-Score with eight examples under the same heuristic.



(a) LeNER-Br

(b) UlyssesNER-Br

Figure 2: Relationship between the number of examples and the F1-score for each corpus.

One possible explanation for these results could be the disparity in the number of training and test examples present in each corpus, as shown in Tables 1 and 2. Specifically, the LeNER-Br corpus contains 28,455 more training examples than the UlyssesNER-Br corpus, providing a broader range for retrieving diverse and informative instances. Similarly, this corpus comprises 6,071 more test examples, enabling a more extensive variety to be evaluated during testing.

Even if there is a disparity in the number of necessary examples to achieve the same F1-Score, it is interesting that they exhibit the same patterns in the heuristic values. In both cases, random sampling had the worst F1-Score of approximately 18% in UlyssesNER-Br and 3.4% in LeNER-Br, indicating that random examples of the use of named entities are unnecessary to provide context for identifying the proper legal entities.

The selection of at least one example per category notably improves the results, particularly when arranging the examples from least similar to most similar, with more similar examples situated near the input sentence. Achieving the best results, with 41%

in UlyssesNER-Br and 43% in LeNER-Br, this is in line with conclusions drawn from previous research (Gupta et al., 2023; Ye et al., 2023; Dong et al., 2022). It is important to emphasize the methodology employed in choosing examples per category, which includes instances with varying degrees of similarity.

This previous selection strategy highlights a noticeable contrast between the most and least similar sentences in the corpus chosen as input examples. In contrast, the selection of *K* similar examples showed that the order of examples does not make much difference in the final result – in fact, both had F1-Score around 51%. However, UlyssesNER-Br requires a more significant number of iterations, as discussed at the beginning of this section. One possible reason results are the same for both ascending and descending orders is that the examples are highly similar and it does not hold that the most similar sentence has a significant effect. This approach indicates that the most suitable examples are selected based on their similarity to the target instance, leading to a better performance outcome for both corpora.

Local Results. Figure 3 demonstrates the entity-wise outcomes in LeNER-Br, while Figure 4 displays the results for UlyssesNER-Br. These graphs depict the influence of each heuristic on individual entities and the number of examples required to achieve specific performance levels.

Upon reviewing the global outcomes, the utilization of random examples did not exhibit any notable enhancement in contextualizing the model, resulting in consistently low F1 scores across all entity types in both corpora. This behavior underscores the significance of selecting pertinent examples for ICL and their potential to enhance the model's adaptability to new tasks. We observed an exception in the random examples concerning the “EVENTO” entity (Figure 4b). This entity presents a comparable performance, although it was the lowest. This may happen due to the relatively small training and validation datasets, as shown in Table 2.

Remarkably, choosing at least one example per category yielded notable results for the ”ORGANIZACAO” entity in UlyssesNER-Br and the ”LOCAL” entity in LeNER-Br (see Figures 4e and 3c). These are unique categories in which the heuristic had similar or better results in a relationship when choosing the top *K* similar sentences. While the outcomes displayed similarities, further investigation might suggest potential improvements in these heuristics that could enhance their effectiveness.

In general, the best heuristic was the selection of the top *K* similar examples. Even though the use of Sabiá with ICL with this heuristic did not provide better results than the state-of-the-art in these corpora (Zanuz and Rigo, 2022; Albuquerque et al., 2022), it is an initial approach that indicates the power of LLMs to extract legal named entities. In the best case, we found three classes in UlyssesNER-Br that obtain more than 50% of F1-Score, being“DATA” with 81%, “LOCAL” with’ 69%, and “PESSOA” with 60%. In LeNER-Br, we also find three classes with F1-Score of over 50%: “PESSOA” with 68%, “ORGANIZACAO” with 53%, and “JURISPRUDENCIA” with 53%.

5.2 Analysing Mispredictions

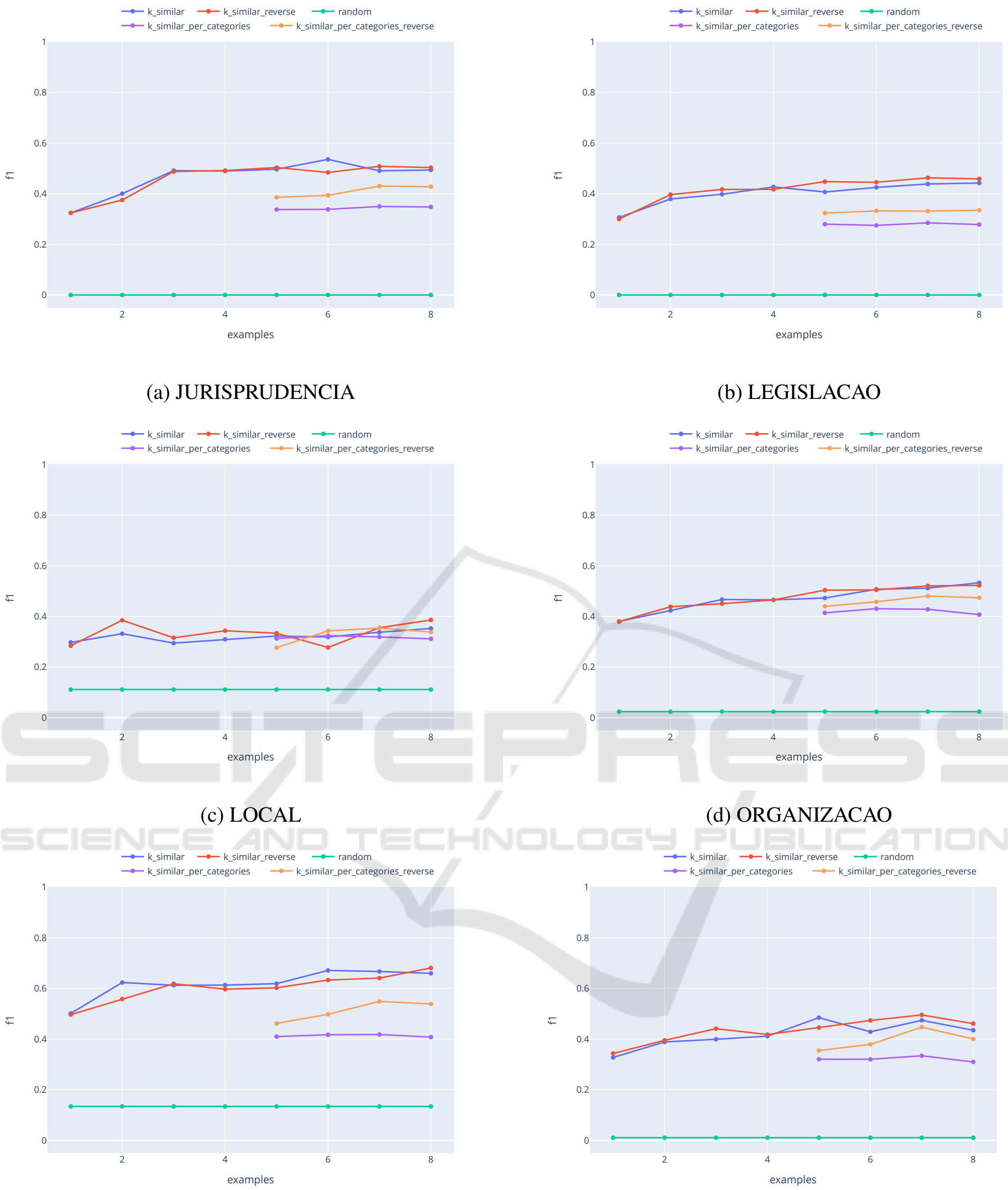
In this section, we discuss misclassification in the LeNER-Br corpus with Sabiá. We present possible explanations without focusing on incomplete classifications of a part of a sentence. Figures 5 and 6 show bar charts with the number of token entities with wrong classifications for each type for the LeNER-Br and UlyssesNER-Br datasets, respectively.

5.2.1 LeNER-Br

“LOCAL”: The model has produced incorrect predictions for this entity, wrongly categorizing terms as “PESSOA” or “ORGANIZACAO”. The prediction as “PESSOA” is the only incorrect classification assigned to *Estado de São Paulo*. However, it is interesting that the four wrong predictions as “ORGANIZACAO” are similar sentences that include the term *Estado do Rio de Janeiro*, which has a potentially ambiguous meaning. For instance, in the following sentence, *Estado do Rio de Janeiro* can be interpreted as a part of the name of the bureau (i.e., “Secretaria de Controle Externo do Estado do Rio de Janeiro”) or as the place where the bureau is located (i.e., “Secretaria de Controle Externo” at the “Estado do Rio de Janeiro”), which can generate different valid classifications for the same term.

> *Entretanto, recebeu o ofício já com prazo assinado de 15 dias para apresentar defesa perante a Secretaria de Controle Externo do Estado do Rio de Janeiro – Secex/RJ, após transcorridos meses (peça 66, p. 6); b) Na Secex/RJ, procurou contato com o Secretário, sem sucesso, tendo conversado com Sérgio Honorato, o qual estranhou o fato de não ter sido ouvido no âmbito da TCE .*

“ORGANIZACAO”: This is the entity most often misclassified by the model, with wrong predictions varying across different types. In contrast to mispredictions of the entity “LOCAL”, the opposite hap-



(a) JURISPRUDENCIA

(b) LEGISLACAO

(c) LOCAL

(d) ORGANIZACAO

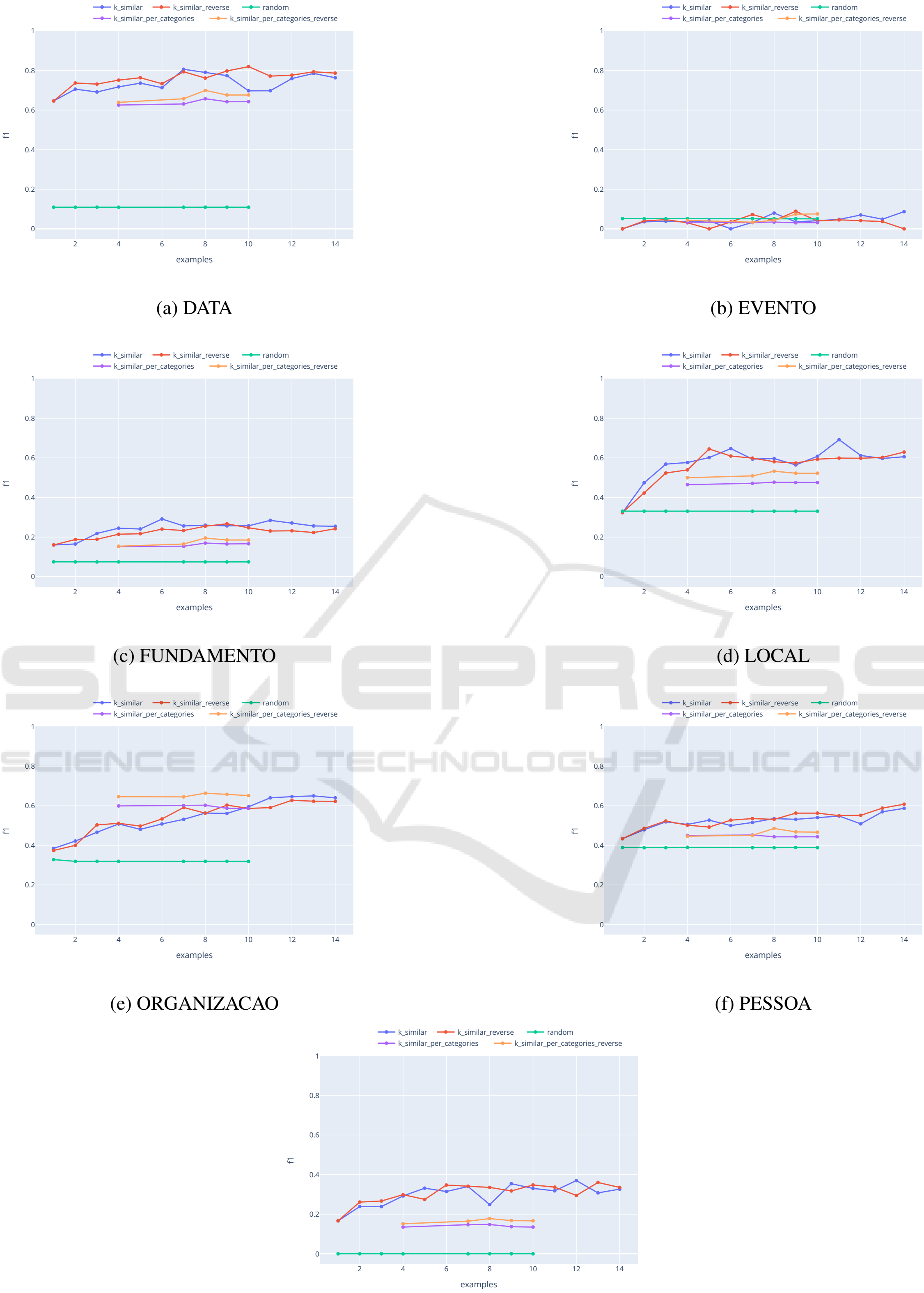
(e) PESSOA

(f) TEMPO

Figure 3: Relationship between the number of examples and the F1-score for each entity in LeNER-Br corpus.

pens now, as an "ORGANIZACAO" is predicted as a "LOCAL". Interestingly, *Rio de Janeiro* again appears four times and with the same ambiguity, as expressed in the following example.

> *Assunto Recurso de Reconsideração interposto por Carlos Aureliano Motta de Souza (ex-Diretor-Geral do Superior Tribunal Militar) contra decisão que julgou suas contas irregulares e o condenou em débito e ao pagamento de multa em razão de irregularidades nas obras de construção do prédio da 1ª Circunscrição Judiciária Militar no Rio de Janeiro.*



(a) DATA

(b) EVENTO

(c) FUNDAMENTO

(d) LOCAL

(e) ORGANIZACAO

(f) PESSOA

(g) PRODUTODELEI

Figure 4: Relationship between the number of examples and the F1-score for each entity in UlyssesNER-Br corpus.

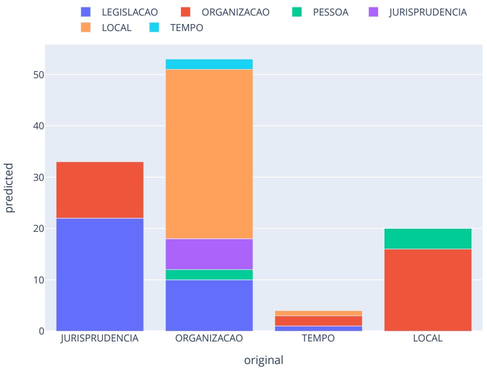


Figure 5: Number of token entities with wrong classifications for each type for LeNER-Br.

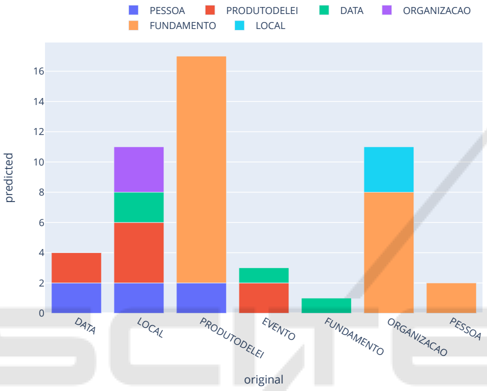


Figure 6: Number of token entities with wrong classifications for each type for UlyssesNER-Br.

Other misclassifications are caused not by ambiguity of meaning, but by the nature of legal text, which has some domain-specific structures and jargon. One example is *Plenário* (i.e., a collegiate court). Even though it should be considered an organization, it sometimes appears together with references to jurisprudence or legislation, separated by a dash (e.g., *Decisão 877/2000 – Plenário* and *Acórdão 2.471/2013 – Plenário*). This can confuse the LLM, which produces the wrong results.

"JURISPRUDENCIA": Terms that should be classified as this entity are often wrongly categorized as "ORGANIZACAO" or "LEGISLACAO". Concerning the terms classified as "ORGANIZACAO", we found a pattern in which only *a part of the part* of the term that includes the organization name is classified. This is what happens with *Tribunal Superior do Trabalho* in the following sentence:

> *Ressalte-se que, estando o v. Acórdão recorrido em sintonia com Orientação Jurisprudencial da C. Corte Superior, tem-se que a sua função uniformizadora já foi cumprida na pacificação da controvérsia, inclusive no que se refere a eventuais violações legais e constitucionais aplicáveis à questão (OJ SDI-I n° 336, do C. Tribunal Superior do Trabalho), não se constatando, outrossim, contrariados outros dispositivos constitucionais não citados no precedente jurisprudencial que embasou o julgado, o que inviabiliza a admissibilidade do apelo também por violações nos termos da alínea "c" , do art. 896, da CLT.*

Misclassifications may also result from similar formatting of terms that refer to things of different concepts. This makes some terms that should be classified as "JURISPRUDENCIA" (*jurisprudence*) be misclassified as "LEGISLACAO" (*legislation*) instead. For instance, both law *§ 3º do art 492 - a da consolidação das leis do trabalho* and jurisprudence *§ 1º-a do artigo 896 da clt* are formatted in exactly the same way despite referring to different concepts, leading to misclassification. The same happens with law *lei nº 11705 de 2008* and jurisprudence *lei nº 68301980 artigo 13 §1º*. The similarity between "JURISPRUDENCIA" and "LEGISLACAO" also makes it particularly difficult to retrieve contextual examples by ICL when the types coexist, as what happens with *Decisão n. 633/1999*, which is classified as "LEGISLACAO" instead of "JURISPRUDENCIA".

"TEMPO" entity has the most misclassification examples. However, even if the classifications were wrong, we could find an annotation error in the corpus, where the LLM could perform the correct classification to *MP/TCU* (The Public Accounting Ministry at the Federal Audit Court) as an "'ORGANIZACAO'.

5.2.2 UlyssesNER-Br

"LOCAL": The model misclassifies terms that should be of this entity type as "ORGANIZACAO", "DATA", "PESSOA", or "PRODUTODELEI". In several cases, however, even if the classifications are technically wrong (i.e., they do not match term annotations in the corpus), they *make semantic sense*. This happens, for instance, with the term *jornal diário catarinense*, which is labeled as a "LOCAL" (i.e., location) in the corpus, but is misclassified by the model as an organization ("ORGANIZACAO"), which is reasonable as the term refers to a newspaper. An analogous case is the term *zona franca de manaus*: it is misclassified as being a "PRODUTODELEI" (i.e., the outcome of a law), which makes semantic sense because it is both an actual place and a public policy (a free economic zone, which is defined by law).

A variation of this phenomenon happens with terms *2000* and *2016*, which, in the context in which they appear (a URL), should have been classified as "LOCAL", but are considered in isolation by the model and misclassified as "DATA" (i.e., date). In addition to these more ambiguous cases, there are also clear errors, such as the term *Brasil*, which is misclassified as a "PESSOA" (i.e., person).

"EVENTO": Three terms that should have been classified as this entity type were misclassified as "DATA" or "PRODUTODELEI". An interesting case is when the model correctly identified "2002" as "DATA" (i.e., a date), but it was actually part of the phrase *eleição presidencial de 2002* (i.e., 2002 presidential election), which should have been an "EVENTO" (i.e., an event). At the same time, the model correctly classified as "EVENTO" the very similar phrase *eleição presidential de 2010* (i.e., 2010 presidential election), which was present in the same context. One possible explanation for this is that the model may have struggled to discern the context due to the presence of dates in different examples and its placement in the middle of the samples. Additionally, the term *Prêmio Nobel* (i.e., Nobel Award) was incorrectly categorized as "PRODUTODELEI".

"FUNDAMENTO": Some terms that should be of this entity type were misclassified as "DATA". As with "EVENTO", the model *does make* correct prediction, but only for a *part* of the phrase instead of the n-gram as a whole. For instance, the term *lei nº 1.043, de 2003* (i.e., law no. 1,043 of 2003) should be categorized as "FUNDAMENTO", but only the year (*2003*) was identified and as a "DATA" (i.e., date) entity.

"PESSOA": Some terms that should be of this type are misclassified as "FUNDAMENTO". An example is *Allan Kardec*, which is a person's name and should therefore be considered a "PESSOA" (i.e., person). Reviewing the contextual examples, we found no clear bias that could be producing this misclassification.

"ORGANIZACAO": Most terms that should have been classified as being of this type are misclassified as "LOCAL" or "FUNDAMENTO". The term *cine rex*, which is classified as a "LOCAL" (i.e., place), is particularly interesting, as it could actually refer to an organization (i.e., considering *cine rex* as a company) or to the location where an exhibit is being held. This term has therefore an intrinsic ambiguity, in that it can be correctly classified as multiple entities.

"PRODUTODELEI": Terms of this type were misclassified as "PESSOA" or "FUNDAMENTO". The term wrongly categorized as "PESSOA" (i.e., person) was *pessoa jurídica*. This is intriguing since it means *legal entity* (e.g., a corporation) and is, therefore, an organization ("ORGANIZACAO"), being thus neither something that is produced by law (i.e., "PRODUTODELEI") nor a person. Terms wrongly classified as "FUNDAMENTO", in turn, seem to be so due to structural similarities to those that are correctly classified as such.

"DATA": Four terms of this type were incorrectly predicted to be either "PESSOA" or "PRODUTODELEI". We found no pattern that could explain these misclassifications.

5.3 Effect of Filtering

Despite providing specific goal labels and their corresponding in-context examples in our prompt, as detailed in Section 3.3, the model occasionally outputs different classifications that were not part of our intended goals but might be valuable in other contexts, such as "ADJETIVO" (adjective), "AÇÃO" (action), and "VERBO" (verb). Additionally, it occasionally produces misspelled entity names, such as "LEGISACAO," "ORGANIZAO," and "PESSA," or with extra punctuation, like "LEGISLACAO;", "LOCAL;" and "ORGANIZACAO;".

While achieving the best F1-Score of 51% in LeNER-Br (as shown in Figure 2a), we observed F1-Score decreased 10% without any filtering and an extra 46 entity types. Similarly, in UlyssesNER-Br, we noted F1-Score decreased 9% and an additional 31 entity types, compared to the best result of 51% (Figure 2b).

By implementing filters to remove punctuation and special characters such as '\n', we observed a decrease of 2% and 3% in F1-Score, and a reduction of 23 and 13 extra entities in LeNER-Br and UlyssesNER-Br, respectively, in comparison with the original result of 51%. Finally, incorporating a similarity filter allowed us to obtain the desired entities as the final outputs, underscoring the capability of Language Model Models (LLMs) to learn from the provided prompts and contexts. However, it also highlights the necessity of post-processing to clean the results and rectify misspellings.

6 CONCLUSION

In this paper, we explored alternatives for legal text classification using In-Context Learning with Sabiá, an LLM fine-tuned for Brazilian Portuguese. To achieve this, we used two corpora of Brazilian legal entities, LeNER-Br and UlyssesNER-Br, which encompass both general entities and those specific to the legal domain. Various retrieval strategies were applied to identify suitable examples in the ICL context. In addition, we employed post-processing techniques to eliminate noise and irrelevant labels from the dataset. In summary, we found that selecting the top K examples is the best heuristic in the general context. Filtering techniques also play an important role in obtaining final results.

We also explored the delved analysis of the model predictions. Along with classification errors, we identified inherent complexities in legal language, which pose challenges for models trained on general domain data. Notably, we observed instances of misannotation that the ICL-enhanced model could classify accurately. Moreover, the model exhibited partial classifications, correctly assigning categories to terms but overlooking the context within which the term operates, leading to a different expected classification. Additionally, our investigation revealed instances of term ambiguities, indicating discrepancies between the predictions and annotations of the model. Despite these discrepancies, both the model and the annotations presented plausible results.

For both corpora, we achieved an F1-Score result of 51% with the best metrics associated with retrieving the most similar examples. These experiments underscored the potential of NER when using LLMs. However, there is a need for further research in this area. Our future work will explore additional heuristics for retrieving relevant documents, experimenting with different prompt templates, and leveraging domain-specific knowledge to enhance predictive accuracy.

ACKNOWLEDGMENTS

This study was partially funded by the Brazilian funding agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

Ahuja, K., Hada, R., Ochieng, M., Jain, P., Diddee, H., Maina, S., Ganu, T., Segal, S., Axmed, M., Bali, K., et al. (2023). Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

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. (2022). Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In *International Conference on Computational Processing of the Portuguese Language*, pages 3–14. Springer.

Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., and Kazi, M. (2023). Transformer models: an introduction and catalog.

Barale, C., Rovatsos, M., and Bhuta, N. (2023). Do language models learn about legal entity types during pretraining? *arXiv preprint arXiv:2310.13092*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Choi, E., Levy, O., Choi, Y., and Zettlemoyer, L. (2018). Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905*.

Cui, L., Wu, Y., Liu, J., Yang, S., and Zhang, Y. (2021). Template-based named entity recognition using bart. *arXiv preprint arXiv:2106.01760*.

Dai, H., Song, Y., and Wang, H. (2021). Ultra-fine entity typing with weak supervision from a masked language model. *arXiv preprint arXiv:2106.04098*.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. (2022). A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Epure, E. V. and Hennequin, R. (2022). Probing pre-trained auto-regressive language models for named entity typing and recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1408–1417.

Fei, Y., Hou, Y., Chen, Z., and Bosselut, A. (2023). Mitigating label biases for in-context learning. *arXiv preprint arXiv:2305.19148*.

Feng, Y., Qiang, J., Li, Y., Yuan, Y., and Zhu, Y. (2023). Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957*.

Gupta, S., Singh, S., and Gardner, M. (2023). Coverage-based example selection for in-context learning. *arXiv preprint arXiv:2305.14907*.

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., Peng, B., Gao, J., and Han, J. (2020). Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978*.

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., and Sayed, W. E. (2023). Mistral 7b.

Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., and Shardlow, M. (2023). Bless: Benchmarking large language models on sentence simplification. *arXiv preprint arXiv:2310.15773*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X. and Qiu, X. (2023). Finding support examples for in-context learning.

Lin, B. Y., Lee, D.-H., Shen, M., Moreno, R., Huang, X., Shiralkar, P., and Ren, X. (2020). Triggerner: Learning with entity triggers as explanations for named entity recognition. *arXiv preprint arXiv:2004.07493*.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2021). What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). 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*, pages 313–323. Springer.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Overwijk, A., Xiong, C., and Callan, J. (2022a). Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362.

Overwijk, A., Xiong, C., Liu, X., VandenBerg, C., and Callan, J. (2022b). Clueweb22: 10 billion web documents with visual and semantic information. *arXiv preprint arXiv:2211.15848*.

Pires, R., Abonizio, H., Rogério, T., and Nogueira, R. (2023a). Sabi\'a: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Pires, R., Abonizio, H., Rogério, T., and Nogueira, R. (2023b). Sabi\'a: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Rubin, O., Herzig, J., and Berant, J. (2021). Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Shen, Y., Tan, Z., Wu, S., Zhang, W., Zhang, R., Xi, Y., Lu, W., and Zhuang, Y. (2023). Promptner: Prompt locating and typing for named entity recognition. *arXiv preprint arXiv:2305.17104*.

Sun, T., Shao, Y., Qian, H., Huang, X., and Qiu, X. (2022). Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Winston, P. H. (1980). Learning and reasoning by analogy. *Communications of the ACM*, 23(12):689–703.

Ye, J., Wu, Z., Feng, J., Yu, T., and Kong, L. (2023). Compositional exemplars for in-context learning. *arXiv preprint arXiv:2302.05698*.

Zanuz, L. and Rigo, S. J. (2022). 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*, pages 219–229. Springer.

Zhang, Y., Feng, S., and Tan, C. (2022). Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*.

Ziyadi, M., Sun, Y., Goswami, A., Huang, J., and Chen, W. (2020). Example-based named entity recognition. *arXiv preprint arXiv:2008.10570*.