Prompt Engineering for Named Entity Extraction from Portuguese Legal Documents

Giovanni Maffeo, Catarina Silva, and Hugo Gonçalo Oliveira
Centre for Informatics and Systems of the University of Coimbra (CISUC)
Department of Informatics Engineering (DEI)
Polo II, Pinhal de Marrocos, 3030-290
Coimbra, Portugal
maffeomedeiros@student.dei.uc.pt
{catarina,hroliv}@dei.uc.pt

Abstract

The growing volume and complexity of legal texts highlight the need for automatic methods capable of extracting structured information from unstructured documents. Motivated by the limited availability and high cost of annotated legal data, this challenge is even more severe for the Portuguese language. This work investigates whether prompt engineering over Large Language Models (LLMs) can effectively support legal Named Entity Recognition (NER) in low-supervision and low-resource settings through In-Context Learning (ICL). Using the LeNER-Br corpus, we evaluate category-specific prompts, different chunking sizes, and prompt engineering strategies. Entity-level evaluation using Exact Match Micro F1 shows that prompt engineering has a stronger impact on performance than other strategies. The best results were obtained with larger models, the 4-bit quantised Qwen-2.5:32B and GPT-5.2, achieving scores of 57.9% and 71.9%, respectively, highlighting the potential of this approach as an alternative to traditional supervised NER pipelines.

1 Introduction

According to Katz et al. (2020), which analyses federal legislation in the United States and Germany, an impressive expansion in both the size and complexity of laws has been observed over the past two and a half decades. The study associates this growth with the modernisation of society, which has given rise to new social behaviours and modes of interaction. As a consequence, legal systems have become denser and more intricate, reflecting their role in regulating an increasingly complex society.

In this context, a larger and more elaborate legal system gives rise to distinct challenges for the judicial system, that is, the institutional framework responsible for adjudicating cases before the courts in accordance with the law. Accordingly, in the Portuguese context, these challenges are clearly observable, with one of the most critical being the prolonged duration of judicial proceedings, to the point that the Portuguese State has been condemned by the European Court of Human Rights (ECtHR) in 75% of the 345 cases against the State (between 1959 and 2018), with the main violation cited being the excessive delay of the Portuguese judicial system[1].

This situation is illustrated by high-profile financial cases such as BES[2], which lasted nearly a decade without the trial beginning and saw several crimes reach the statute of limitations. Similarly, the BANIF[3] case, where victims denounced delays and unmet compensation expectations. Furthermore, the *Destaque Estatístico Anual - 2024*[4], reports a total of 589,507 pending cases as of 30 June 2025 and anticipates a stagnation in the reduction of this figure in the coming years.

This situation of delays in resolving legal cases have far-reaching implications, from court congestion and increasing legal costs for the parties involved, to public perception of justice and fairness. Altogether, this scenario highlights an urgent need for innovative solutions capable of streamlining judicial processes and improving access to justice.

In this context of innovation, techniques based on Natural Language Processing (NLP) have emerged as effective solutions capable of handling

[1] https://expresso.pt/sociedade/2020-01-15-Atrasos-preconceitos-interferencia-desproporcionada.-Os-casos-em-que-o-Estado-portugues-falhou. Accessed: March 2026.

[2] https://www.rtp.pt/noticias/pais/besges-tribunal-declara-prescricao-de-11-crimes-do-processo_a1604586. Accessed: March 2026.

[3] https://www.rtp.pt/noticias/economia/lesados-do-banif-pedem-reuniao-a-montenegro-e-denunciam-promessas-nao-cumpridas_n1706076. Accessed: March 2026.

[4] https://webapi-estatisticas.justica.gov.pt/api/public/ContentMedias?id=bf5c32c8-79cb-406c-9e7c-33bdd01acb34&lang=PT. Accessed: March 2026.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1092–1097

the large volume of unstructured, natural-language data produced daily in the legal domain (Chalkidis et al., 2020; Zheng et al., 2021). Within this landscape, one particularly relevant NLP task is Named Entity Recognition (NER), which consists of automatically identifying and classifying mentions of entities, such as persons, organisations, locations and temporal expressions, in free text (Mota et al., 2010). Naturally, when applied to a specialised domain, NER must also recognise domain-specific entities, which are often not trivial for non-experts to identify. For example, in the legal domain, entities such as *legislation* and *legal cases* require specialised knowledge to annotate correctly, as their meaning and structure differ substantially from general-purpose named entities.

It is important to highlight that NER is not only a core NLP task on its own, but also a fundamental building block for a wide range of downstream applications. These include information extraction, question answering, text summarisation, relation extraction, knowledge base construction and legal document retrieval (Li et al., 2022).

In this direction, works have been developed to support NER in the Portuguese legal domain (Luz de Araujo et al., 2018; Albuquerque et al., 2022; Brito et al., 2023). In recent years, these approaches have typically relied on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), but their effectiveness is constrained by the high cost and limited availability of large, high-quality annotated legal corpora.

Motivated by these limitations, recent advances in decoder-based Large Language Models (LLMs) have introduced a promising alternative. Beyond their conversational capabilities, those models have demonstrated strong performance across a wide range of NLP tasks, including machine translation and question answering, even under zero-shot and few-shot conditions (Yu et al., 2023; Ma, 2023).

Given this motivation and the fact that legal Portuguese is a niche domain with only a few publicly available annotated resources, which significantly constrains the development of supervised models, this work investigates the potential of LLMs for legal NER through prompt engineering and In-Context Learning (ICL), representing a low-supervision setting since no or very few annotated examples are necessary.

This paper makes two main contributions. First, it investigates whether the prompt-based framework originally designed for clinical NER (Zhang et al., 2023) can be effectively adapted to the Portuguese legal domain, with slight modifications that account for domain-specific characteristics. Second, it extends one of the future directions suggested in the original work by evaluating the performance of quantised and open-source LLMs on the same NER task, as the prior focused exclusively on commercial models.

The rest of this paper is structured as follows. Section 2 introduces our methodology, including the research questions and experimental setup. Section 3 presents our experimental results and discussion. Finally, Section 4 concludes the paper.

2 Methodology

Experiments were guided by a set of research questions and corresponding hypotheses, detailed in Section 2.2. All research questions were investigated through NER annotation experiments conducted on the LeNER-Br corpus (Luz de Araujo et al., 2018), a publicly available dataset specifically designed for NER in the Portuguese legal domain. Each research question applies an isolated adaptation of the baseline configuration in order to identify the most effective NER techniques for this dataset.

Despite the relevance of the work and the support provided by the dataset's author, we conducted additional experiments using the CDJUR-BR corpus (Brito et al., 2023) to assess the reproducibility of our results across different legal corpora. Although it contains different entity types, it is substantially larger, which could facilitate the application of supervised approaches. However, during the annotation analysis we encountered labeling rules that we were unable to clearly explain, which we attributed to inconsistencies in the dataset. These issues ultimately led us to discontinue its use in the experimental setup. One example is the expression "Ministério Público", which in some documents was annotated as *Person* and in others not annotated as an entity, even when appearing in identical contexts, such as "O Ministério Público Estadual, por meio da Promotora de Justiça signatária, no uso de suas atribuições legais...".

2.1 Baseline Experiment

The baseline and other experiments were carried out using a setup detailed in Section 2.3. Our baseline model was the *Qwen 3* LLM, with 8.2B parameters, a context window of 40,960 tokens,

an embedding dimension of 4,096 and Q4_K_M quantization. The decoding parameters included a temperature set to 0, top_k of 20 and top_p of 0. The model operated on document chunks with a maximum length of 900 characters. This chunk size was chosen because legal documents tend to be lengthy and, if processed in full, would exceed the model's context window. Finally, the models were accessed from different sources depending on the experiment. Quantized models were executed locally using Ollama[5] and through a private API provided by an external service, while OpenAI models were accessed via the OpenAI API[6].

2.2 Research Questions

Below, we present the research questions and their associated experimental design.

1. **Research Question 1 (RQ1):** *Does the usage of category-specific prompts improve entity annotation?* This experiment explores the usage of **specific-prompt** strategies, in which each entity type is annotated in a separate prompt execution.

2. **Research Question 2 (RQ2):** *How does the size of the context window affect NER performance?* This experiment investigates the effect of **chunking** on entity annotation, aiming to understand how the effective input context window size influences performance. Three configurations were evaluated: chunks with a maximum size of 300 characters (H2a), 600 (H2b) and sentence-based chunking (H2c).

3. **Research Question 3 (RQ3):** *To what extent do prompt engineering techniques improve NER performance in the legal domain?* This experiment evaluates the impact of **prompt engineering** techniques. We adapt the framework proposed in Zhang et al. (2023) to the Portuguese legal domain. The following conditions were tested:

   - Baseline prompts only (Baseline or H3a);
   - Annotation guideline-based prompts (H3a + H3b);
   - Annotation guideline-based prompts with no-context examples (H3a + H3b + H3c);
   - Error analysis-based instructions (H3a + H3b + H3c + H3d);
   - One-shot prompting (H3a + H3b + H3c + H3d + H3e);
   - Five-shot prompting (H3a + H3b + H3c + H3d + H3e + H3f).

   Figure 1 presents a representative snippet of the most complete prompt configuration, highlighting the sections incorporated with each hypothesis (H3a-H3f). The baseline task description defines the annotation objective and output format (H3a), entity definitions and no-context examples correspond to H3b and H3c. Example directives derived from manual error analysis correspond to H3d, and contextualized input-output example pairs correspond to H3e and H3f. Note that the inclusion of H3c is an adaptation beyond what was explored in the original clinical NER study. The source code and the complete prompt templates are available in the project repository[7].

4. **Research Question 4 (RQ4):** *Do selected high-performing strategies generalise to unseen data and alternative models?* This experiment evaluates whether the high-performing strategy generalises to an unseen test set and whether its results can be further improved when applied to alternative LLM.

H3a - Baseline
Your task is to generate an HTML version of an input text, marking specific entities related to the legal domain. The entities are ...

H3b and H3c - Guidelines and no-context examples
Legislation is defined as expressions that identify legal norms, such as laws and codes (e.g., "Lei nº 8.666/93", ...).

H3d - Error analysis
Words such as "súmula", "acórdão" or "jurisprudência", when used in a generic way and without reference to a specific decision, should not be annotated as jurisprudence ...

H3e and H3f - Few-shot prompting
Example input: O MINISTÉRIO PÚBLICO FEDERAL ...
Example output: O `<span class="organizacao">` MINISTÉRIO PÚBLICO FEDERAL `</span>` ...

Figure 1: Snippet of the prompt configuration (H3a-H3f).

[5] https://ollama.com
[6] https://developers.openai.com/api/docs
[7] https://github.com/giovannimaffeo/prompt-engineering-portuguese-legal-ner

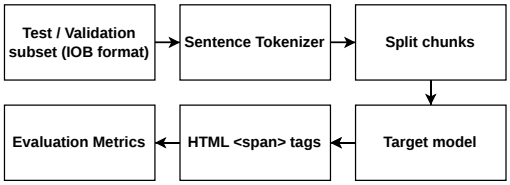

Figure 2: Overview of the experimental setup used in this study.

2.3 Experiment Setup

The experimental setup follows the workflow illustrated in Figure 2. For Research Questions RQ1 to RQ3, all evaluations were performed on the `validation` split of the LeNER-BR dataset (Luz de Araujo et al., 2018), while RQ4 was conducted on the `test` split in order to assess model generalisation.

To apply the chunking strategies required for RQ2, we processed the original LeNER-BR files in IOB (Inside–Outside–Beginning) format and segmented the text using the `PunktSentenceTokenizer` from NLTK[8]. Sentences were grouped into the largest possible sequences without exceeding the maximum character length defined for each experiment. This ensured that document segments remained consistent with the chunk-size constraints evaluated in our study.

After preprocessing, the resulting chunks were passed to the target LLM for annotation. Following Zhang et al. (2023), we reformulated the NER task as a text generation problem by instructing the models to highlight named entities using HTML `<span>` tags, with each tag containing the corresponding entity type. This format allows the model's output to be easily converted back into the traditional IOB tagging scheme.

Below is an illustration of the annotation format used in our experiments:

Example input

O MINISTÉRIO PÚBLICO FEDERAL, por meio do Procurador CARLOS HENRIQUE ALVES, propôs ação com base no art.61 , §1º , da Lei nº 9.430/96

Example output

O MINISTÉRIO PÚBLICO FEDERAL, por meio do Procurador CARLOS HENRIQUE ALVES, propôs ação com base no art.61 , §1º , da Lei nº 9.430/96

[8] https://www.nltk.org/

Hypothesis	Model	Data	F1
Base (H3a)	Qwen-3:8B	valid	0.374
H1	Qwen-3:8B	valid	0.394
H2a	Qwen-3:8B	valid	0.401
H2b	Qwen-3:8B	valid	0.395
H2c	Qwen-3:8B	valid	0.411
H3b	Qwen-3:8B	valid	0.379
H3c	Qwen-3:8B	valid	0.418
H3d	Qwen-3:8B	valid	0.480
H3e	Qwen-3:8B	valid	0.527
H3f	Qwen-3:8B	valid	0.463
H4a	Qwen-3:8B	test	0.531
H4b	LLaMA-3.1:8B	test	0.398
H4c	Gemma-3:12B	test	0.514
H4d	Ensemble	test	0.544
H4e	Qwen-2.5:32B	test	0.579
H4f	GPT-5.2	test	0.719

Table 1: Results hypotheses using Exact Match F1.

2.4 Evaluation Metrics

For evaluation, an entity-level approach was used, as described in Jurafsky and Martin (2023, Chap. 9, p. 214). However, we encountred similar limitations as Zhang et al. (2023). As noted by the authors, GPT-based models face challenges when identifying correct entity boundaries under exact-match evaluation. For example, while the gold annotation may contain only the specific legal reference (e.g., "Art. 10"), the model may output a longer and contextually valid span such as "Art. 10 do Código Civil". Although the expanded phrase is semantically accurate, it violates the strict boundary required by the annotation schema. Unlike BERT-based models, which are fine-tuned directly on annotated spans with well-defined limits, GPT-based models are trained on a broader and more diverse corpus, which can reduce their ability to adhere strictly to exact entity boundaries. As a result, boundary sensitivity under exact-match evaluation constitutes an inherent limitation of prompt-based approaches to NER.

3 Results

The experimental results are presented in Table 1, following the research questions and hypotheses defined in Section 2.2.

Regarding **RQ1**, we observed that annotating each entity type in a separate prompt execution led to improvements over the baseline configuration. For **RQ2**, following the intuition that asking the model to annotate smaller chunks should be simpler than annotating larger text chunks, annotating single sentences (H2c) exhibited the best results, followed by a maximum length of 300 characters (H2a), 600 characters (H2b) and 900 characters

(H3a or baseline).

Regarding **RQ3**, our results show that prompt engineering is a key factor in improving performance, confirming that a prompt-based framework originally designed for clinical NER (Zhang et al., 2023) can be effectively adapted to the Portuguese legal domain. While the inclusion of guidelines and examples led to incremental gains over the baseline, the most substantial improvement was achieved by explicitly incorporating common error patterns into the prompt (H3d), which increased the Micro F1 score from 41.8% to 48.0%. Additionally, the one-shot strategy (H3e) achieved the best overall results, indicating that increasing the number of examples does not necessarily lead to better performance, as observed with the five-shot configuration (H3f).

Regarding **RQ4**, we evaluated the proposed approach on the test subset using the best-performing prompt configuration (H3e) and observed that the results generalised well to unseen data. Among the smaller models (H4a, H4b and H4c), all using 4-bit quantisation, Qwen-3:8B achieved the best individual performance. The ensemble strategy combining these three models (H4d) outperformed the weaker individual models, but did not outperform larger architectures (H4e and H4f). Finally, when scaling to larger architectures, the 4-bit quantised Qwen-2.5:32B and the commercial GPT-5.2 achieved Micro F1 scores of 57.9% (H4e) and 71.9% (H4f), respectively. These results demonstrate that prompt-engineered LLMs can achieve competitive performance for legal NER in Portuguese under low-supervision and low-resource settings, requiring no or very few manually annotated training examples. When compared to the results reported by Luz de Araujo et al. (2018), where a LSTM-CRF model achieved an F1 score of 92.53% through supervised fine-tuning, our best results approach this benchmark. This highlights the potential of prompt-based approaches as an alternative to traditional supervised NER pipelines.

To further investigate the model's performance, we conducted a distribution error analysis on a random sample of 5 chunks from the outputs generated by GPT-5.2 (H4f), which is also detailed in the project repository. The error statistics, presented in Figure 3, reveal that boundary errors constitute the most critical failure mode (50%, $n = 7$), followed by missing entities (35.7%, $n = 5$), incorrect extractions (14.3%, $n = 2$) and incorrect entity type (0%, $n = 0$). Notably, boundary errors

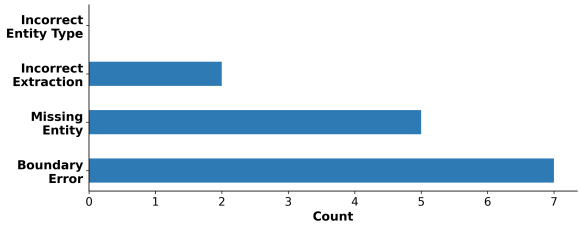

Figure 3: Error distribution analysis (random sample of 5 chunks).

are particularly detrimental in exact-match NER evaluation as they generate dual penalties. Each boundary mismatch simultaneously produces one false negative (i.e., the complete gold entity was not predicted correctly) and one false positive (i.e., the partially predicted entity does not exist in the gold standard). For instance, when the model predicted `006.010/2000-4` instead of the gold entity `Processo 006.010/2000-4` for the entity type *legislation*, this single boundary error contributed both to reduced recall and precision, thereby having a multiplicative impact on F1 score degradation. This explains why boundary errors, despite representing only half of the observed failures, have disproportionate effects on overall performance metrics. The boundary errors primarily occurred in complex multi-token legal entities, such as incomplete statutory citations and partial organizational names (e.g., predicting `TST` when the gold annotation was `C. TST` or predicting `DEJT 12.3.2010` instead of `12.3.2010`). These cases indicate that prompt-engineered LLMs struggle to determine precise entity boundaries in highly specialized legal terminology. Thus, this may explain the lack of performance in terms of the F1 score compared to Luz de Araujo et al. (2018), whose approach relies on fine-tuning with annotated entities that have clearly defined boundaries, whereas LLMs are trained on broader and more diverse corpora.

4 Final remarks

In conclusion, our results indicate that ICL can support NER in the Portuguese legal domain without supervised fine-tuning, with prompt engineering having a stronger impact than task decomposition or context reduction. Overall, it represents a viable alternative to traditional supervised NER approaches, particularly in low-supervision settings.

Acknowledgements

This work was partially supported by the European Fund for Regional Development (FEDER), through

the Program Centro2030, in the scope of project nº 14439, "AI4JURIS – Assistente jurídico de última geração". This work was also funded by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 – Centre for Informatics and Systems of the University of Coimbra (CISUC).

References

Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. Silva, Diego Vitório, and André C. P. L. F. de Carvalho. 2022. Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language (PROPOR)*. Springer.

Maurício Brito, Vládia Pinheiro, Vasco Furtado, João Araújo Monteiro Neto, Francisco das Chagas Jucá Bomfim, André C. F. Costa, Raquel Silveira, and Nilsiton Aragão. 2023. Cdjur-br: Uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. Sociedade Brasileira de Computação.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Nikos Aletras. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing*. Draft Edition, available online.

Daniel Martin Katz, Michael James Bommarito, and Josh Blackman. 2020. Complex societies and the growth of the law. *arXiv preprint arXiv:2005.07646*.

Junjie Li, Aixin Sun, Jie Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A dataset for named entity recognition in brazilian legal text. In *Proceedings of the International Conference on the Computational Processing of Portuguese (PROPOR)*.

Chao Ma. 2023. Prompt engineering and calibration for zero-shot commonsense reasoning. *arXiv preprint*, arXiv:2304.06962.

Cristina Mota, Diana Santos, Luís Miguel Cabral, and Nuno Silva. 2010. Segundo harem: Avaliação de reconhecimento de entidades mencionadas em português. In *Actas do Encontro Nacional da Associação Portuguesa de Linguística*.

Fei Yu, Linda Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Hao Zhang, Jingyan Wang, Shuai Yu, and 1 others. 2023. Improving large language models for clinical named entity recognition via prompt engineering. *arXiv preprint*.

Lucy Zheng, Neel Guha, Patrick Ivy, Daniel E. Ho, and Peter Henderson. 2021. When does pretraining help? assessing self-supervised learning for law and the case of casehold. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.