Evaluating Prompting Strategies for Clinical Information Extraction in Portuguese using LangExtract

Bruno Rogério S. dos Santos[1,2], **Eugénio Ribeiro**[2,3], **Leonardo S. Rocha**[1]

[1]LAGIC, State University of Ceará
Fortaleza – CE – Brazil

[2]ISTAR, University Institute of Lisbon (ISCTE-IUL)
Lisbon – Portugal

[3]INESC-ID
Lisbon – Portugal

bruno.rogerio@aluno.uece.br

Eugenio.Ribeiro@iscte-iul.pt, leonardo.sampaio@uece.br

Abstract. *This work presents an initial systematic evaluation of LangExtract, an LLM-based extraction library, applied to LusoClin, a dataset of clinical notes in European Portuguese. The study compares zero-shot and few-shot prompt strategies for NER, using prompts adapted to the clinical domain and evaluating performance with Precision, Recall, and F1-score in Strict, Exact, Partial Match and Entity Type. When compared to a fine-tuned model and a rule-based model, the results show that LangExtract achieves performance close to that of mmBERT (F1-score up to 0.79), suggesting that LLM-based extraction via LangExtract can be a viable alternative for information extraction without the need to train on large volumes of data.*

1. Introduction

Information Extraction (IE) transforms unstructured text into structured data, and in clinical settings, IE supports tasks like diagnosis, treatment, and research by parsing narratives such as nursing or radiology notes. Despite its value, clinical IE is hindered by scarce annotated data, linguistic complexity, and the need for domain expertise [Sivarajkumar et al. 2024]; challenges that are amplified for languages other than English, including Portuguese, as data volumes keep growing [Bender 2019].

LLMs have reshaped IE by replacing task-specific training with prompting strategies like zero-shot and few-shot learning, allowing broader generalization with less supervision. Yet their performance is highly sensitive to prompt design and example quality: few-shot prompting tends to boost accuracy on complex tasks but demands careful example selection and tuning, while zero-shot methods risk lower accuracy and inconsistent structured outputs [Al Nazi et al. 2025]. For Portuguese specifically, standardized benchmarks and evaluation frameworks remain scarce, complicating comparisons across approaches, and linguistic variability plus domain-specific terminology make prompt design even harder.

Motivated by these gaps, the article evaluates clinical IE in Portuguese by comparing zero-shot and few-shot prompting strategies, assessing the LangExtract library (using GPT-4o-mini and Gemini 3.1 Flash Lite) on the LusoClin

[Félix and Teixeira Lopes 2025] clinical corpus against rule-based and mmBERT [Marone et al. 2025] baselines, and discussing broader challenges in prompt design, evaluation, and generalization for Portuguese LLM-based IE. The paper is organized into six sections covering related work, methodology, results, discussion, and conclusions with future directions.

2. Related Work

2.1. Clinical Information Extraction in Portuguese

Works on IE for Portuguese clinical and medical texts remains considerably smaller than the corresponding literature for English, but important progress has been made in recent years. Among IE subtasks, Named Entity Recognition (NER) is the most mature area [Da Rocha et al. 2022].

For Relation Extraction (RE), there is progress, but it remains more limited than NER. Compared to English, Portuguese RE studies are even smaller, often integrated into broader streams rather than presented as stand-alone reference tasks. This reveals a significant gap: although entity recognition models are improving, there is less work that systematically models the relationships between diagnoses, medications, procedures, and outcomes in Portuguese health texts. A relevant effort in this direction is the development of two clinical datasets in Brazilian Portuguese, DGO and SemClinBr, composed of annotated entities and relationships extracted from medical records. The first dataset includes 1,622 documents, 146,769 entities, and 111,716 relations, and supports models that achieve an accuracy of 78.24% for entity and relation extraction [Da Silva et al. 2023]. The second dataset includes 1,000 documents, 65,117 entities, and 11,263 relations, and supports a variety of NLP tasks [Oliveira et al. 2022]. Despite the size of both datasets, there is still no set of common and fixed rules to evaluate, in a comparable way, RE systems in clinical texts written in Portuguese.

In Event Extraction (EE), the gap is even greater. The extraction of clinical events seeks to identify clinically relevant occurrences, such as hospitalizations, surgeries, onset of symptoms, medication administration, discharge or disease progression. There are still no widely adopted clinical EE benchmarks in Portuguese comparable to tasks shared in English. More generally for Portuguese, a study [Cunha et al. 2023] proposed a quality control-based event extraction framework and introduced ACE-2005-PT, a Portuguese version of the ACE-2005 corpus, achieving new baselines for extracting triggers and arguments [Cunha et al. 2024]. Although not clinical specific, this work is highly relevant because it demonstrates that EE capabilities in Portuguese can be created through translation and alignment strategies. The clinical domain still lacks equivalent corpora focused on hospital events and temporal medical narratives.

2.2. LLM-based Information Extraction in Clinical Domain

Studies on IE in general have used LLMs. One example is Sabiá-2, a family of LLMs trained on Portuguese texts. Its performance on question answering was evaluated in several exams, including medical exams [Almeida et al. 2024]. Sabiá-2 Medium equals or surpasses the performance of GPT-4 in 23 of 64 exams and surpasses GPT-3.5 in 58 of 64 exams, at a cost per token ten times lower than that of GPT-4. Despite its strength

on knowledge benchmarks, Sabiá-2 has not been evaluated on structured clinical IE tasks like NER or relation extraction.

Recent work has begun to establish more solid foundations for extracting clinical information in Portuguese. The latest benchmark study [Anjos de Almeida et al. 2026] compared fine-tuned versions of BioBERTpt, BERTimbau, ModernBERT, and mmBERT with LLMs such as GPT-5 and Gemini-2.5, using the public SemClinBr [Oliveira et al. 2022] corpus and a private breast-cancer dataset. Among the systems evaluated, mmBERT-base achieved the best overall performance, reaching a micro-F1 of approximately 0.76 and outperforming the other models. For European Portuguese, Medi-Albertina [Nunes et al. 2024] also showed clear gains in NER, improving F1 by more than 30% over BioBERTpt and 15% over BERTimbau in a single-model setting. In addition, one of the few LLM-based clinical IE frameworks developed for European Portuguese, applied to IPO clinical texts, reported a precision of 88% with Zephyr-7b-beta, surpassing Med42-70b, which reached 70% precision [Leite 2024].

When compared to the works mentioned here, this article uses a dataset in European Portuguese, LusoClin [Félix and Teixeira Lopes 2025], which contains more than 54,000 clinical notes, a dataset significantly larger in quantity compared to the datasets cited in this work, and entirely focused on clinical texts. Although NER is the main subtask associated with this dataset, it is also possible to extract clinical relationships and events. And, last but not least, the use of LLMs through the LangExtract library allows for cost reduction related to training and processing for the NER task, as will be shown in section 4.

3. Methodology

LangExtract is a prompt-based library for LLM-driven IE that enables users to extract structured information from unstructured text by defining extraction schemas and optional examples, without the need to train dedicated models. It enforces consistent structured outputs (exportable to JSONL) through controlled generation, supporting parallel processing of full documents from URLs and interactive visualization of large entity sets [Goel 2025]. While offering notable advantages such as cross-domain flexibility and reduced dependence on annotated data, it also presents challenges regarding output consistency, prompt sensitivity, and evaluation.

To evaluate LangExtract in the clinical domain, we developed an experiment with clinical texts in European Portuguese, specifically in the LusoClin dataset [Félix and Teixeira Lopes 2025], following a controlled extraction flow, from input normalization to structured output generation and post-processing. Prompts were developed to guide the model in extracting clinically relevant entities, using domain-specific instructions and examples to reduce ambiguity and improve consistency. The corpus is synthetic and was generated from real clinical data (SemClinBr [Oliveira et al. 2022] and CLINpt [Lopes et al. 2019]) using fine-tuning in Llama 3.2 with QLoRA, and prompt engineering. LusoClin includes abbreviations, incomplete phrases, and region-specific terminology, making the task more challenging and closer to a real-world scenario.

3.1. Domain and Corpora

LusoClin [Félix and Teixeira Lopes 2025] is a public dataset consisting of fully synthetic clinical notes in European Portuguese built to be a new benchmark. These notes

were carefully validated to preserve clinical realism and consistency. The corpus contains 54,518 notes, distributed across four categories: 14,253 admission notes, 13,561 ambulatory notes, 9,032 discharge summaries, and 17,672 nursing notes, all excluding real patient data[1]. For the experiments, the notes were split into training examples and a separate test set. The entity annotation was performed using MediAlbertina PT-PT [Nunes et al. 2024], a BERT-based model for the clinical domain. This choice shifts the emphasis of the study to evaluating how closely the proposed approach matches the annotation behavior of that model. Therefore, the test set, after the application of LangExtract, was used as a reference, and all extracted entities were aligned with the annotations according to the chosen matching criteria. The evaluation was performed on the entire portion intended for testing, specifically 250 clinical records from each category.

3.2. Extraction Pipeline

LusoClin clinical notes are processed by LangExtract, which identifies and organizes relevant clinical entities from unstructured text using a prompt that guides extraction. The notes follow an XML-like structure containing the text, entity ID, tag, start and end offsets, and the extracted entity text. LangExtract produces structured JSONL output, making the data machine-readable and suitable for evaluation against reference scores after recalculating offsets. Figure 1 represents the LangExtract pipeline in this work.

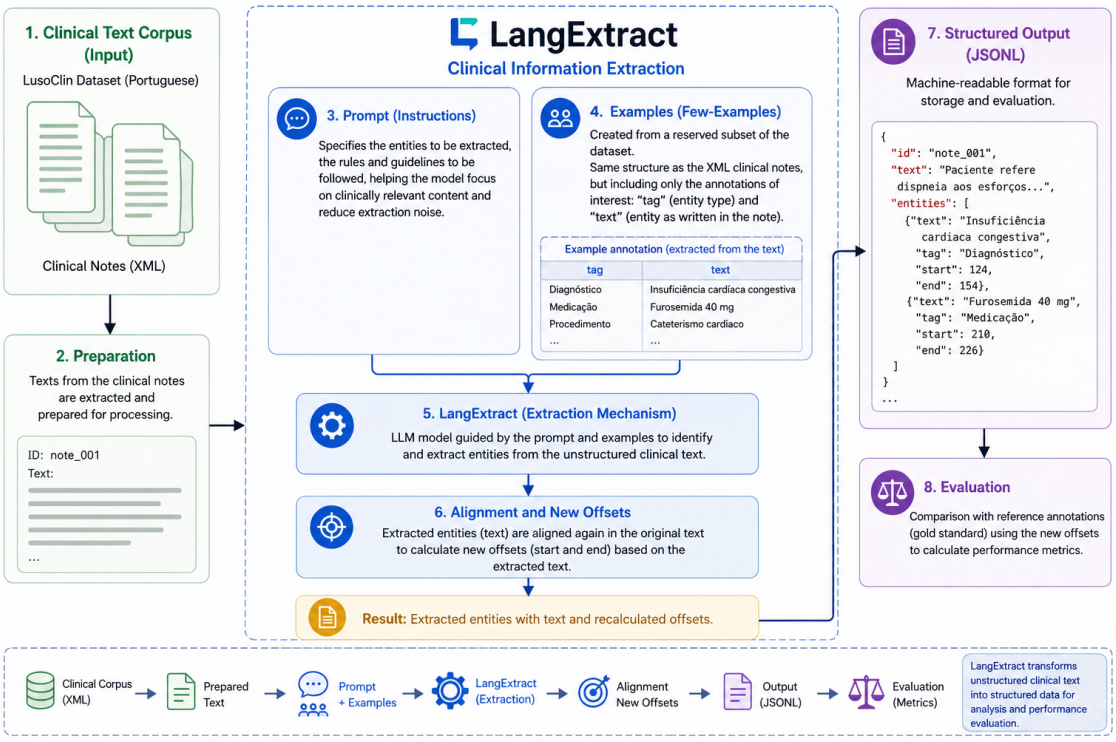


Figure 1. LangExtract Pipeline.

3.3. Prompt Design

Two prompts written in European Portuguese were evaluated under zero-shot and few-shot settings. The first was a simple prompt designed to assess the contribution of con-

[1] https://rdm.inesctec.pt/ca/dataset/cs-2025-005

textual examples, instructing the model to extract clinical entities in document order, preserve the original text, select the most complete and specific spans, avoid paraphrases and generic terms, and identify the eight LusoClin entity categories: **Diagnosis, Symptom, Medical Procedure, Medication, Dosage, Vital Sign, Result, and Progress**. The second was a more detailed prompt developed from preliminary observations, providing explicit extraction rules and descriptions of each entity type to improve span precision and reduce dependence on examples, particularly in the zero-shot setting. Thus, the prompts enabled comparison of a simple, example-oriented strategy with a more structured and descriptive strategy across both prompting conditions. The two prompts used are publicly available in a GitHub repository, which will be cited in the next subsection.

To evaluate the effectiveness of the proposed extraction pipeline, we compared its performance with two reference models. The first is a deterministic rule-based model, built with regular expressions and geographic dictionaries. The second is mmBERT [Marone et al. 2025], a Transformers-based model fine-tuned on LusoClin. This comparison allows us to situate the proposed LLM-based pipeline in relation to both traditional rule-based methods and next-generation tuned language models.

3.4. Implementation Details

For the LangExtract experiments, we used OpenAI GPT-4o-mini (gpt-4o-mini-2024-07-18)[2] [Hurst et al. 2024] and Google Gemini 3.1 Flash Lite[3], accessed through their respective cloud APIs. Both models were selected for cost-efficient, high-volume text processing[4], with Gemini recommended by the LangExtract documentation [Goel 2025]. LangExtract received clinical notes and task instructions, with or without contextual examples, and returned entities in JSONL format. Since LangExtract uses generative decoder-based LLMs, no traditional training was required; instead, contextual learning was evaluated using k examples with $k = 2$, 4, 8, 16, and 32, following [Wang et al. 2025], and in addition to one-shot prompting. Three runs were performed for each configuration, with examples randomly sampled from the corresponding category and converted into *ExampleData* objects containing the original text and reference annotations. Random seeds ensured reproducibility, and temperature was set to 0. Inference was performed entirely through the OpenAI and Google APIs, requiring no local GPU resources; the total cost was $77.84.

The rule-based model, on the other hand, used patterns for recurrent clinical expressions, dosages, measurements (e.g., numerical values followed by units such as mg, mL, and mcg), laboratory values, medications, and other entity types, complemented by dictionaries for the eight dataset labels. Although computationally inexpensive and effective for recurring patterns, its coverage is limited for linguistic variation, unknown terminology, abbreviations, and context-dependent mentions. LusoClin preprocessing retained notes without entities for mmBERT as valid negative examples, handled overlapping entities, excluded entity-free notes from LangExtract examples, and normalized MediAlbertina annotations by standardizing labels, text, and offsets while removing invalid and duplicate annotations. For mmBERT, 80% of the non-test notes were used for

[2] https://developers.openai.com/api/docs/models/gpt-4o-mini

[3] https://deepmind.google/models/gemini/flash-lite/

[4] https://openai.com/pt-BR/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

training and 20% for validation; word-level labels were aligned with subword tokens, with non-initial and special tokens assigned -100. Fine-tuning used a learning rate of 2×10^{-5}, batch size of 16, up to 10 epochs, and early stopping with a patience of three epochs based on validation F1-score.

The rule-based and mmBERT experiments were conducted on a Windows 11 system with an AMD Ryzen 7 5700X CPU, NVIDIA GeForce RTX 3070 8GB GPU, and 32GB RAM. The scripts, prompts, and experimental configurations are publicly available in the GitHub repository [5]

3.5. Evaluation Method

To evaluate performance, we used standard metrics for the IE task: precision, recall, and F1 score. Precision measures the proportion of extracted entities that are correct, indicating how many of the system's predictions are relevant. Recall measures the proportion of reference entities that were successfully retrieved, indicating the completeness of the extraction. The F1 score is the harmonic mean of precision and recall and provides a single, balanced measure of overall performance.

To compare predicted entities with reference annotations, we considered different matching criteria. In strict match, a predicted entity is considered correct only when both its range and its label exactly match the reference annotation. This is the most rigorous evaluation setting and rewards precise boundary detection. In exact match, a predicted entity is considered correct only if its range exactly matches the reference annotation. This setting measures the ability to correctly locate spans, ignoring the entity label. In partial match, a prediction is considered partially correct when there is overlap between the predicted and reference ranges, even if the boundaries are not identical. This criterion is useful in clinical texts, where entity boundaries may be ambiguous or difficult to consistently delineate. Furthermore, we assessed entity type agreement to verify if the predicted label corresponds to the correct semantic category, regardless of the range quality. This allows for a more detailed analysis of whether errors arise from boundary detection, type classification, or both.

4. Results

The results are presented in two ways: for the few-shot scenario, graphs report the mean and standard deviation of F1-score across three runs to assess stability, while mean F1-scores alone are used to distinguish zero-shot from few-shot approaches and compare LangExtract with other models. The results show that few-shot performance depends primarily on the number of examples rather than prompt formulation. Using 8 examples already approaches the best performance, while 16 examples yields F1-scores practically equivalent to 32 examples, particularly for Partial Match and Entity Type, as can be observed in Figures 2 and 3. In contrast, zero-shot consistently achieved the lowest scores, whereas even one example produced substantial improvements, especially for Partial Match and Entity Type. Overall, few-shot achieved the highest F1-scores across admission, outpatient, discharge, and nursing notes, with the largest gains in the more demanding Strict Match and Exact Match evaluations.

[5] https://github.com/brusimoes/LusoClin_Experiments

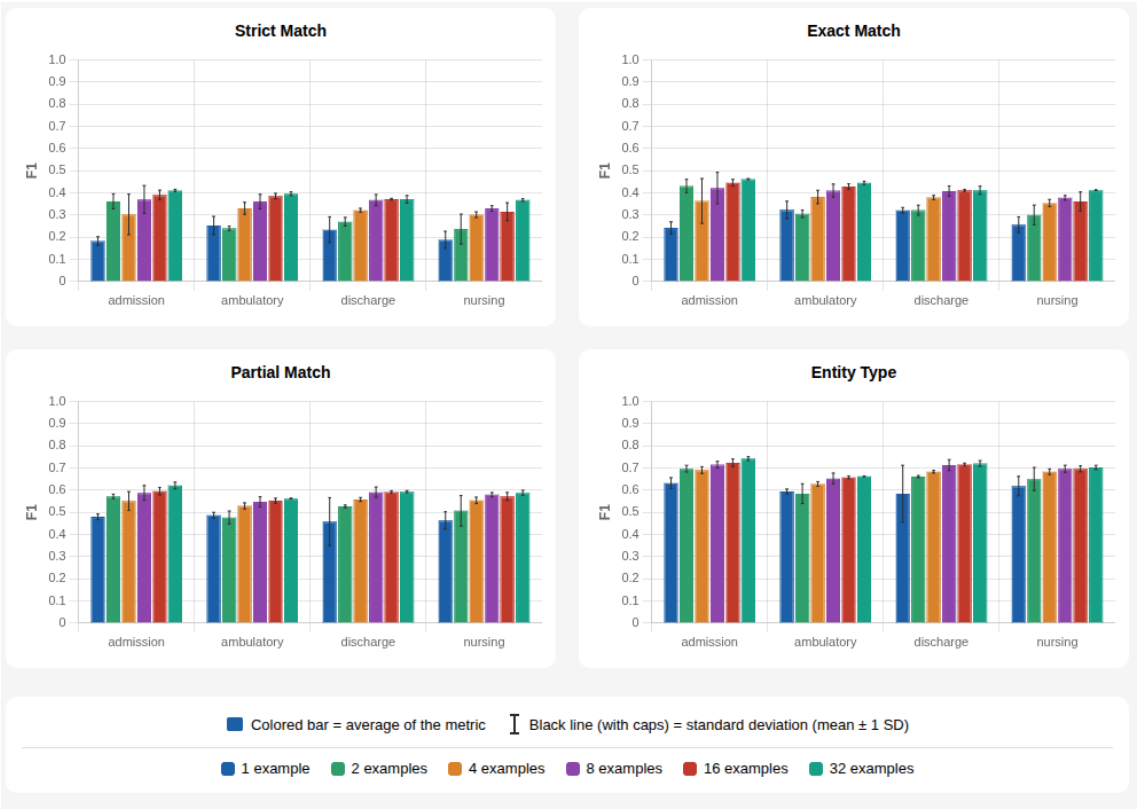


Figure 2. LangExtract (GPT-4o-mini) using Simple Prompt with examples.

For LangExtract, the best GPT-4o-mini configuration was the simple prompt with 16 examples; this was compared with Gemini 3.1 Flash Lite using the same setting, with the simple prompt selected over the slightly better descriptive prompt due to its lower prompt size and associated cost. As shown in Figure 5, Gemini achieved the best LangExtract results across all match criteria, reaching an F1-score of approximately 0.79. However, among all evaluated approaches, fine-tuned mmBERT achieved the highest overall performance, with an F1-score of up to 0.87, outperforming both LangExtract and the rule-based model, as shown in the Figure 6.

For LangExtract, the best GPT-4o-mini configuration was the simple prompt with 16 examples; this was compared with Gemini 3.1 Flash Lite using the same setting, with the simple prompt selected over the slightly better descriptive prompt due to its lower prompt size and associated cost. As shown in Figure 5, Gemini achieved the best LangExtract results across all match criteria, reaching an F1-score of approximately 0.79. However, among all evaluated approaches, fine-tuned mmBERT achieved the highest overall performance, with an F1-score of up to 0.87, outperforming both LangExtract and the rule-based model, as shown in the Figure 6.

5. Discussion and Limitations

The results show that few-shot prompting consistently outperforms zero-shot extraction, with the best F1-scores obtained using a simple prompt and 16 examples. The improvements are particularly evident for Partial Match and Entity Type, indicating that demonstrations help identify clinical concepts and entity types, although precise boundary de-

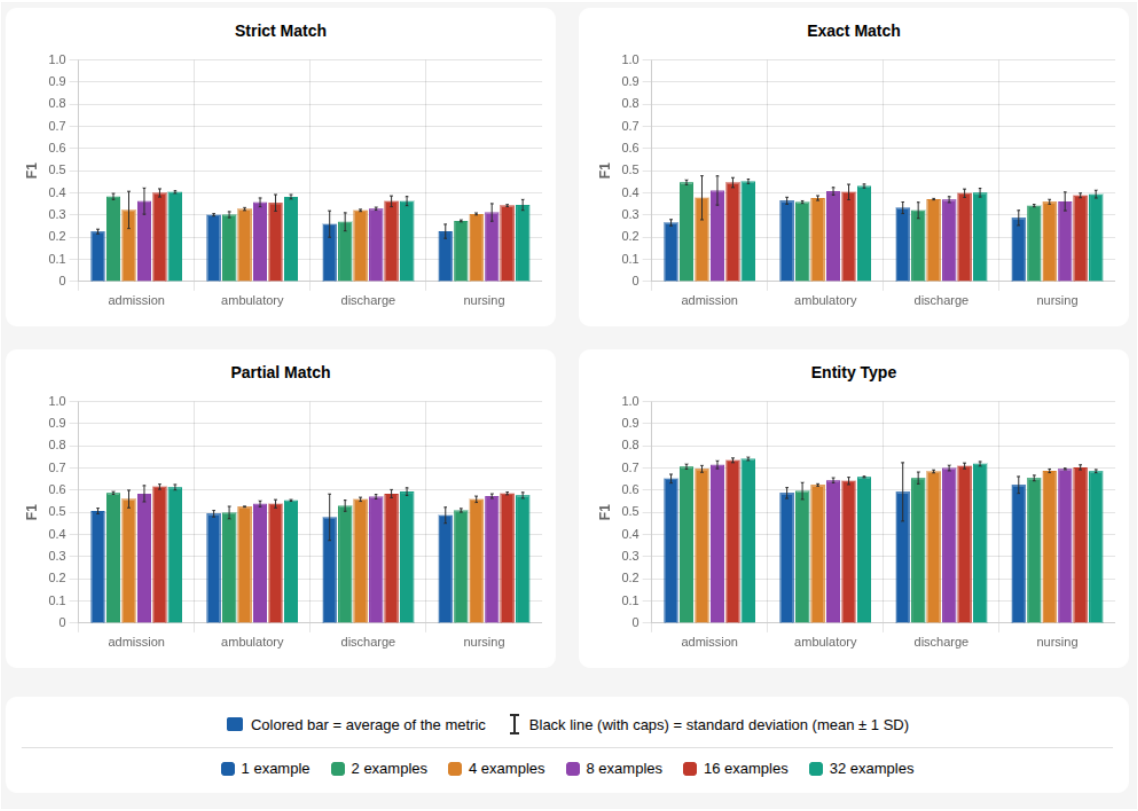


Figure 3. LangExtract (GPT-4o-mini) using Descriptive Prompt with examples.



Figure 4. Results of mean F1-score values from LangExtract using GPT-4o-mini.

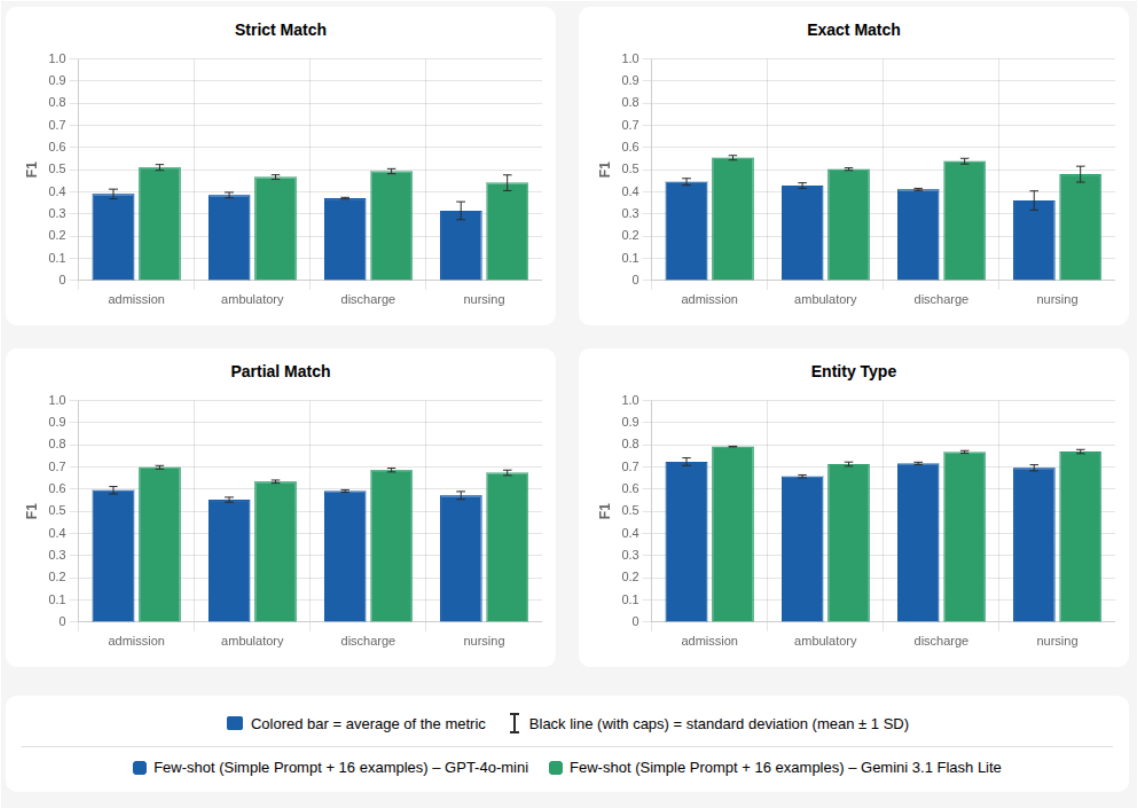


Figure 5. F1-score results with LangExtract (GPT-4o-mini x Gemini 3.1 Flash Lite).

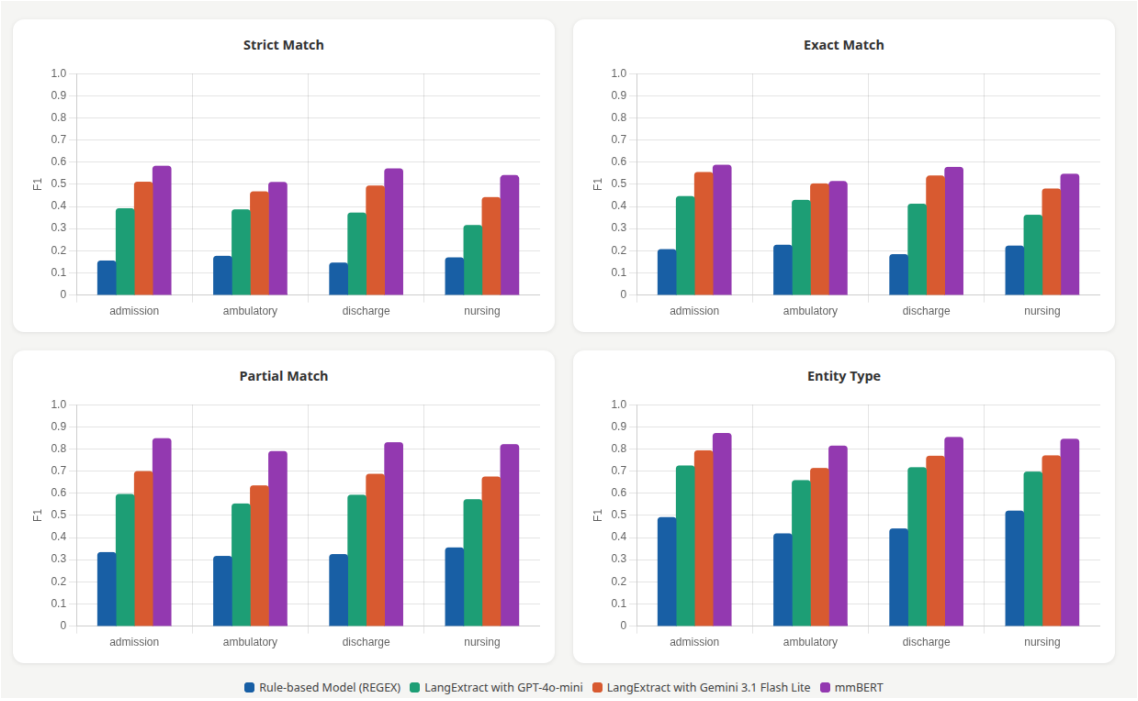


Figure 6. Results of mean F1-score values.

tection remains challenging for short expressions, abbreviations, and multi-word entities. Performance is also sensitive to prompt formulation and example selection, suggesting that model behavior depends strongly on contextual examples rather than on a fully robust understanding of the task. The higher performance observed with only two examples in admission notes (Figure 2 and Figure 3) may therefore result from particularly representative examples, the heterogeneous writing styles and entity patterns of these notes, and the greater importance of example quality and coverage than of quantity alone.

This study has some limitations, the most important being that the evaluation measures how well the proposed approaches reproduce the annotations automatically generated by MediAlbertina, rather than comparing their performance to a manually validated gold standard. Annotation errors, including incorrect entity types and notes without annotations, were observed in MediAlbertina outputs. Moreover, the evaluation may structurally favor mmBERT because it is fine-tuned on MediAlbertina-derived annotations and evaluated against annotations produced by the same system and related BERT-based modeling paradigm. Thus, the results may reflect the ability to reproduce MediAlbertina's patterns and errors rather than true clinical entity extraction performance. A manually validated benchmark reviewed by clinical experts would provide a more reliable and independent evaluation, and the reported results should therefore be interpreted as performance relative to MediAlbertina-derived references rather than definitive evidence of model superiority.

Another limitation concerns the evaluation of Gemini 3.1 Flash: the best-performing configuration identified for GPT-4o-mini was directly transferred to Gemini 3.1 Flash, which may not represent the ideal configuration for the latter model. Although a more extensive exploration implies higher computational and financial costs, a minimal additional experiment using the largest number of evaluated examples (k=32) would be valuable to determine if the current Gemini results are potentially suboptimal and if increasing the number of examples can improve its performance. Overall, although the use of a small number of examples with few cases has proven beneficial in several situations, the threshold detection errors and sensitivity to selected demonstrations confirm that extracting clinical information in Portuguese remains a challenging and open research problem.

6. Conclusion

This work evaluated LangExtract on the LusoClin dataset for clinical IE in European Portuguese using LLMs, achieving an F1 score of up to 0.79, a result close to that obtained by mmBERT with fine-tuning, which was 0.87. These results demonstrate that LangExtract can compete with domain-tuned models, making them particularly relevant for resource-limited scenarios where annotated corpora are scarce and the development of task-specific systems is restricted.

Unlike conventional LLMs that require specific training for domain adaptation, LangExtract allows for rapid adaptation through prompt engineering, reducing the cost and time associated with building specialized IE systems. This is especially significant for Portuguese, where NLP (Natural Language Processing) resources remain more limited than in English, positioning it as a promising alternative to accelerate both research and large-scale clinical implementation while maintaining robust and competitive perfor-

mance.

Future work should address two important limitations of the current study. First, constructing a manually validated evaluation subset of LusoClin should be prioritized. Such a subset would provide an independent reference standard against which the automatically generated annotations could be assessed, reducing the dependence on the original dataset annotations and enabling a more reliable evaluation of entity extraction quality. Despite requiring manual annotation effort, this is a feasible extension and would substantially strengthen the empirical validity and scientific contribution of the study.

Second, the best-performing configuration identified for GPT-4o-mini was directly transferred to Gemini 3.1 Flash, which may not represent the optimal configuration for the latter model. Although a more extensive exploration would incur higher computational and monetary costs, a minimal additional experiment using the largest evaluated number of examples (k=32) would be valuable to determine whether the current Gemini results are potentially suboptimal and whether increasing the number of examples can improve its performance. Scalability at a broader level will also be evaluated across larger and more complex datasets, reinforcing LangExtract's potential as an essential tool for information extraction as LLM-based methods continue to evolve.

Acknowledgments

We would like to express our sincere gratitude to CAPES for the financial support provided through the Doctoral Sandwich Program Abroad (PDSE), which made this research possible. We also extend our appreciation to ISCTE for the institutional and logistical support offered during the research stay, providing the necessary infrastructure and academic environment for the development of this work. This work was supported by Portuguese national funds through *Fundação para a Ciência e a Tecnologia, I.P.* (FCT) under projects UID/50021/2025 (DOI: 10.54499/UID/50021/2025), UID/PRR/50021/2025 (DOI: 10.54499/UID/PRR/50021/2025), and UID/04466/2025 (DOI: 10.54499/UID/04466/2025).

References

Al Nazi, Z., Hossain, M. R., and Al Mamun, F. (2025). Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, 10:100124.

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*.

Anjos de Almeida, V., Saorin da Silva, S., Chire, J., Vicenzi, L., Borges, N. H., Kociolek, H., Miriã de Castro Rocha, S., Nassif Gomes, F., Ferreira, J. C., Marques, O., et al. (2026). Clinical named entity recognition in the portuguese language: a benchmark of modern bert models and llms. *arXiv preprint arXiv:22603.26510*.

Bender, E. M. (2019). The bender rule: On naming the languages we study and why it matters. The Gradient.

Cunha, L. F., Campos, R., and Jorge, A. (2023). Event extraction for portuguese: a qa-driven approach using ace-2005. In *EPIA Conference on Artificial Intelligence*, pages 402–414. Springer.

Cunha, L. F., Silvano, P., Campos, R., and Jorge, A. (2024). Ace-2005-pt: Corpus for event extraction in portuguese. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 661–666.

Da Rocha, N. C., Barbosa, A. M. P., Schnr, Y. O., Machado-Rugolo, J., de Andrade, L. G. M., Corrente, J. E., and de Arruda Silveira, L. V. (2022). Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1):11.

Da Silva, D. P., da Rosa Fröhlich, W., De Mello, B. H., Vieira, R., and Rigo, S. J. (2023). Exploring named entity recognition and relation extraction for ontology and medical records integration. *Informatics in medicine unlocked*, 43:101381.

Félix, D. and Teixeira Lopes, C. (2025). Dataset of synthetic clinical notes in european portuguese generated using an open-source large language model, along with prompting and evaluation data [data set]. inesc tec.

Goel, A. (2025). LangExtract: Llm-powered structured information extraction from text with source grounding. `https://github.com/google/langextract`.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Leite, M. A. A. (2024). Ontology-based extraction and structuring of narrative elements from clinical texts. Master's thesis, Universidade do Porto (Portugal).

Lopes, F., Teixeira, C., and Oliveira, H. G. (2019). Contributions to clinical named entity recognition in portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233.

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., and Van Durme, B. (2025). mmbert: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

Nunes, M., Boné, J., Ferreira, J. C., Chaves, P., and Elvas, L. B. (2024). Medialbertina: An european portuguese medical language model. *Computers in Biology and Medicine*, 182:109233.

Oliveira, L. E. S. e., Peters, A. C., Da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., and Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G., and Guo, C. (2025). Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275.