Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting

Marco Naguib, Xavier Tannier, Aurélie Névéol
Université Paris-Saclay, CNRS, LISN, Orsay, France
Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, Limics, Paris, France
firstname.lastname@lisn.upsaclay.fr
xavier.tannier@sorbonne-universite.fr

Abstract

Large language models (LLMs) have become the preferred solution for many natural language processing tasks. In low-resource environments such as specialized domains, their few-shot capabilities are expected to deliver high performance. Named Entity Recognition (NER) is a critical task in information extraction that is not covered in recent LLM benchmarks. There is a need for better understanding the performance of LLMs for NER in a variety of settings including languages other than English. This study aims to evaluate generative LLMs, employed through prompt engineering, for few-shot clinical NER. We compare 13 auto-regressive models using prompting and 16 masked models using fine-tuning on 14 NER datasets covering English, French and Spanish. While prompt-based auto-regressive models achieve competitive F1 for general NER, they are outperformed within the clinical domain by lighter biLSTM-CRF taggers based on masked models. Additionally, masked models exhibit lower environmental impact compared to auto-regressive models. Findings are consistent across the three languages studied, which suggests that LLM prompting is not yet suited for NER production in the clinical domain.

1 Introduction

Electronic Health Records (EHR) are rich sources of clinical information (Demner-Fushman et al., 2009), which often appear in unstructured text only (Escudié et al., 2017). Efficiently extracting information from EHRs into a more structured form can help advance clinical research, public health surveillance and clinical decision support (Wang et al., 2018).

Named Entity Recognition (NER) is a critical primary step in information extraction that aims to identify and categorize mentions of relevant entities in text. In the context of clinical information extraction, these can be mentions of clinical entities such as disorders or drugs. Extracting these entities can significantly enhance concept normalization (Cho et al., 2017; Wajsbürt et al., 2021; Sung et al., 2022) as well as facilitate interpreting patient profiling and phenotyping (Gérardin et al., 2022). Clinical NER is widely regarded as a challenging problem : clinical entities are often jargon or ambiguous, and clinical texts have a nonstandard phrasal structure (Luo et al., 2020; Leaman et al., 2015). Additionally, the sensitive nature of EHRs results in a lack of publicly available clinical corpora, which are often restrictively licensed and predominantly available in English. Moreover, the annotation of clinical NER data demands substantial domain expertise, rendering such campaigns both costly and time-intensive (Luo et al., 2020; Névéol et al., 2014; Doğan et al., 2014; Báez et al., 2020). Additionally, due to the diversity of clinical cases, data annotated for one biomedical application might not necessarily be helpful for another. Consequently, there is a critical need for data-efficient clinical NER, specifically few-shot NER approaches.

Large Language Models (LLMs) - specifically, causal, generative models - have demonstrated significant promise in few-shot learning across a wide variety of tasks, including text classification, machine translation, and question answering (Brown et al., 2020; Radford et al., 2019). This supports our main research question: Can few-shot learning performance of LLMs transfer to the task of named entity recognition?

This study aims to evaluate generative LLMs, employed through prompt engineering, for few-shot clinical NER from the perspective of F1 performance and environmental impact.

Challenges of few-shot NER We identify methodological challenges with the evaluation of few-shot prompting for NER.

First, there is no standard, widely adopted manner of prompting LLMs for NER tasks (Li et al., 2022; Shen et al., 2023; Wang et al., 2023b; Ker-

Findings of the Association for Computational Linguistics: EACL 2024, pages 6829–6852

aghel et al., 2024), resulting in significant challenges for reproducibility and variations in results that are difficult to interpret. Second, many efforts towards "few-shot learning" with LLMs design prompts based on their performance on large held-out validation datasets (Brown et al., 2020; Tam et al., 2021; Radford et al., 2021; Qin and Eisner, 2021), which is not consistent with a few-shot setup. In addition, in-context learning performance is shown to depend greatly on the prompt structure: a small change in task phrasing, the examples presented, the order of examples, or the tagging format can affect the performance (Zhao et al., 2021; Lu et al., 2022; Min et al., 2022). Therefore, making these choices assuming large annotated validation dataset leads to performances that are shown (Perez et al., 2021) to be over-optimistic and impossible to find in a real few-shot setting. Third, Zaghir et al. (2024) show that the majority of recent studies employing prompt engineering in medical applications lack a non-prompt-related baseline, such as fine-tuned BERT-like Masked Language Models (MLMs), which complicates the accurate assessment of the performance of these LLMs. Finally, most of these studies are mainly concentrated on English, and based on GPT, which is mainly trained on English, (Jimenez Gutierrez et al., 2022; Wang et al., 2023b; Ashok and Lipton, 2023; Hu et al., 2023b; Zaghir et al., 2024), limiting the generalizability of evaluations to other languages. The contributions of this work are as follows:

1. We describe a systematic algorithm for creating and optimizing prompts for NER, bringing a particular attention to tagging prompts (Wang et al., 2023b), a novel NER prompting technique that recently showed particular promise (García-Barragán et al., 2024; Magron et al., 2024).

2. We evaluate our algorithm in a true few-shot setting, by allowing prompt optimization only on the few annotated instances through cross-validation. 14 NER tasks were evaluated, spanning 6 general-domain datasets and 8 clinical datasets, with a focus on English, French and Spanish.

3. We compare this approach, applied to 13 generative LLMs, to the standard fine-tuning approach, applied to 16 MLMs, both in terms of performance and environmental impact. We provide our code at github.com/marconaguib/autoregressive_ner

2 Few-shot & clinical NER

Few-shot NER with pre-trained MLMs Leveraging MLMs for NER usually involves using them as encoders, and training a linear projection to map vector representations into an NER tagging of the sentence, while jointly fine-tuning the parameters of the language model itself for the downstream task of NER (Devlin et al., 2019). This approach has been widely studied (Liu, 2019; Petroni et al., 2020; Joshi et al., 2020; Schweter and Akbik, 2021). Wajsbürt (2021) propose a similar architecture, enhanced with an entity decoder, to iteratively predict entity spans in the input, allowing the model to detect nested entities.

Few-shot NER can be performed by simply training such systems with the limited data available. Other approaches have been proposed to leverage MLMs specifically in few-shot setting. Namely, metric learning (Fritzler et al., 2019; Yang and Katiyar, 2020; Huang et al., 2021a) proposes to train systems to instead learn a metric over the output space. New instances can then be classified based on the distance separating them from other labeled instances. Label encoding (Aly et al., 2021; Ma et al., 2022a; Hou et al., 2020) suggests, instead, to leverage label names or textual label descriptions and encode them along with the instances in order to better tag them.

Few-shot NER with generative LLMs Recently, prompt construction has gained interest in the community (Brown et al., 2020; Liu et al., 2023). While most related work focused on studying prompt formulation and exploring better-performing prompt structures (Wei et al., 2022; Ashok and Lipton, 2023; Vilar et al., 2023; Wang et al., 2023b) also known as "prompt engineering", other work proposed continuous optimization of the prompt through prompt tuning (Ma et al., 2022b; Layegh et al., 2023; Hu et al., 2023a), usually reporting marginal improvements over baselines.

There is no standard, widely adopted manner of building NER prompts (Liu et al., 2023). In fact, NER associates to each instance a set of spans, each of which having a type. This structured nature of the prediction make it hard to find an intuitive but efficient manner to prompt a language model for NER, that adapts well to all contexts.

For instance, the main practice is to use separate prompts for different entity types (Li et al., 2020;

Liu et al., 2022; Chen et al., 2023a). This choice seems well-suited when the task is interested in a handful of types of entities (typically 5-10). When interested in less entity types, a single prompt can be used for detecting all entities (Ashok and Lipton, 2023). On the other hand, if there is more entity types, it could be interesting to enumerate every possible span in the input sentence and let the model predict the entity type of the span, if any (Cui et al., 2021). This method, on the inverse, is impractical for long inputs.

We identify three strategies for prompting LLMs. Constrained prompting attempts to better formulate the NER task by constraining the generation to fill in specific hand-crafted templates, usually adapted to MLMs (Cui et al., 2021; Shen et al., 2023; Ye et al., 2023; Schick and Schütze, 2021). Listing prompts consist in simply making the language model predict the entities in a list (Ashok and Lipton, 2023). Tagging prompts were studied more recently by Wang et al. (2023b). They make the language model surround entity mentions with special tags.

Few-shot clinical NER MLM-based few-shot NER has also been explored in the biomedical domain (Ge et al., 2023). Metric leaning (Yang and Katiyar, 2020) and label encoding (Aly et al., 2021; Ma et al., 2022a) have been explored, as well as other approaches such as active learning (Kormilitzin et al., 2021), supervised pretraining (Huang et al., 2021b) and prompt-based learning (Lee et al., 2022; Chen et al., 2023b; Cui et al., 2024).

Few studies have focused on LLM-prompting-based few-shot clinical NER. In Hu et al. (2024), GPT-3 and ChatGPT are evaluated on the 2010 i2b2/VA task (Uzuner et al., 2011) in a zero-shot context. In Jimenez Gutierrez et al. (2022), GPT-3 is evaluated on a set of biomedical information extractions tasks including the NCBI-Disease Doğan et al. (2014).

On languages other than English,Meoni et al. (2023) use InstructGPT-3 is used to build multilingual training corpora to train smaller models, and in Ateia and Kruschwitz (2023), ChatGPT is evaluated on an NER challenge focused on extracting medical procedures in Spanish.

Another interesting direction is partly fine-tuning (Liao et al., 2023) a general-domain LLM on clinical text (Han et al., 2023; Toma et al., 2023; Yang et al., 2024), and prompting the resulting LLM.

3 Named Entity Recognition Experiments

3.1 Evaluation tasks

We use 14 publicly-available NER datasets (described in the next section) to compare prompted causal models to fine-tuned masked language models in few shot settings. For each study language, we selected two out-domain datasets and two or three in-domain datasets, aiming to use comparable resources (same genre, tagset, annotation guidelines) across languages whenever possible. We use official training, validation, and test subsets when available; otherwise, we apply an 80%-10%-10% split for training, validation, and testing.

3.1.1 General-domain evaluation datasets

WikiNER (Nothman et al., 2013) is a multilingual *silver-standard* annotated NER dataset. It consists of a late-2010 snapshot of Wikipedia in nine languages. Hyperlinks referring to persons, locations or organizations were automatically annotated. We use the English, French and Spanish versions of this dataset.

CoNLL-2002 (Tjong Kim Sang, 2002) and **CoNLL-2003** (Tjong Kim Sang and De Meulder, 2003) are two manually-annotated multilingual NER dataset released as a part of CoNLL shared tasks. Mentions of persons, locations, organizations and miscellaneous entities are annotated. We use the Spanish data of the 2002 version, which is a collection of news wire articles made available by the Spanish EFE News Agency, released in May 2000. We use the English data of the 2003 version, which consists of Reuters news stories between 1996 and 1997.

Quaero French Press (Grouin et al., 2011) is a manually annotated corpus of about 100 hours of speech transcribed from French speaking radio broadcast. This corpus was used in the 2011 Quaero named entity evaluation campaign. It comprises annotations for 5 entity types further divided into 32 subtypes. Our experiments relied on the five entity types: persons, locations, organizations, functions, and facilities.

3.1.2 Clinical evaluation datasets

E3C (Magnini et al., 2021) is a European multilingual corpus (Italian, English, French, Spanish, and Basque) of semantically annotated clinical narratives. The texts are collected from multiple publicly-available sources such as abstracts extracted from CC-licensed journals. We use the

gold standard material available from the English, French and Spanish versions of this dataset. The clinical narratives are annotated with 6 entity types : actors, body parts, events, RMLs (measurements and test results) and clinical entities.

The **n2c2-2019** (Luo et al., 2020) shared task focuses on medical concept normalization. It uses the MCN corpus developed by (Luo et al., 2019), often referred to as the n2c2-2019 dataset. It includes de-identified discharge summaries from the Partners HealthCare and Beth Israel Deaconess Medical Center. In order to convert the medical concept normalization task into an NER task, we use the annotated Concept Unique Identifiers (CUIs) to map each mention to the corresponding UMLS semantic group (Lindberg et al., 1993; McCray et al., 2001).

The **NCBI-Disease** (Doğan et al., 2014) corpus gathers 793 PubMed abstracts where mentions of diseases are annotated in four types depending on their syntax : Specific Diseases (e.g. *diastrophic dysplasia*), Disease Classes (e.g. *an autosomal recessive disease*), Composite Mentions (e.g. *colorectal, endometrial, and ovarian cancers*), and Modifiers (e.g. *C7-deficient*) .

QuaeroFrenchMed (Névéol et al., 2014) consists of two text sources that we treat separately. The first part, **EMEA** is a collection of 13 patient information leaflets on marketed drugs from the European Medicines Agency (EMEA). The second part, **MEDLINE**, consist of 2,500 titles of research articles indexed in the MEDLINE database[1]. The two parts are annotated with 10 entity types corresponding to UMLS semantic groups.

The Chilean Waiting List (Báez et al., 2020) corpus consists of 900 de-identified referrals for several specialty consultations in Spanish from the waiting list in Chilean public hospitals, manually annotated with 10 entity types : abbreviations, body parts, clinical findings, diagnostic procedure, diseases, family members, laboratory or test results, laboratory procedures, medications, procedures, signs or symptoms and therapeutic procedures. It can be noted that these types can be redundant (e.g. all diagnostic procedures are also annotated as procedures).

3.1.3 Few-shot learning set-up

We simulate the few-shot context by only providing the models with a few annotated examples that can be used in training, prompting and validation. No additional examples are made available. In this study, we choose to mainly focus on $k = 100$ sentences, which corresponds to one to two hours of annotation in the clinical domain (Névéol et al., 2014; Campillos et al., 2018). We use a fixed random seed p to choose k examples among all those available in the corpus. In Section A.1.1, we discuss the effect of the choice of k and of p.

Additionally, we train the best-performing language models with the entire training dataset to provide a skyline comparison, i.e. performance of the models outside the few-shot setting.

3.2 Language Models

Table 5 (appendix A.2) presents an overview of the language models used in our study. English is covered in all causal language models, which is not the case for Spanish and French. Except for mBERT and XLM-RoBERTa, masked language models cover only one of our study languages.

3.3 NER with MLMs

Although MLMs have been adapted to few-shot learning in architectures suited for low-resource contexts (see section 2), we compare LLM prompting to the simple and standard MLM usage without any further adaptation for few-shot learning, as an accessible and easily-reproducible baseline.

We use NLStruct (Wajsbürt, 2021), an open-source Python library [2] that implements the standard fine-tuning approach. NLStruct uses the representations provided by the language model to encode the input, then employs a bidirectional LSTM decoder and a CRF to iteratively predict the entities present in the encoded input, as described by Gérardin et al. (2022). This approach allows NLStruct to effectively handle nested entities, which are prevalent in some of the study corpora.

We train the model for 20 epochs on 80% of the data and use the remaining held-out 20% for early stopping.

3.4 NER with generative LLM prompting

Our experiments prompt models to tag entities in the input sentence, instead of listing them. This choice is supported by further experiments reported in Section A.1.2. The upper part of Figure 1 shows a sample tagging prompt, highlighting sections in the prompt that guided the design of features for prompt phrasing.

[1] http://pubmed.ncbi.nlm.nih.gov/

[2] https://github.com/percevalw/nlstruct

Prompt features We describe below the nine optional features that control the phrasing of the prompt, as well as the criteria for selecting the few-shot examples featured in the prompt.

1. **Prompt language**: By default, we prompt all language models in English, as it is the most ubiquitous language in all of their training corpora. This feature allows the model to be prompted in French or Spanish, to align the prompt language with that of the test sentence.

2. **Additional sentences**: By default, we present 5 annotated sentences in the prompts. This feature presents 5 additional sentences (i.e., 10 sentences in total). Section A.1.3 discusses adding more demonstations to the prompt.

3. **Self verification**: By default, we select the 5 closest sentences to the test sentence in terms of TF-IDF distance, among the training set. The mentions tagged by the model are then considered to be the model's final predictions. This feature selects instead the 5 sentences featuring the most entities of the target type and features them in an initial prompt. Intuitively, this prompt results in higher recall and lower precision. A second "self-verification" prompt is then used over the model's initial predictions in order to filter out the false positives. A sample self-verification prompt is shown in the bottom part of Figure 1.

   The number of demonstrations follows that of the main prompt.

4. **Taggers**: By default, we follow (Wang et al., 2023b) prompting the model to surround mentions with @@ and ##. This feature prompts it to surround mentions with quotes « and » instead.

5. **Address a specialist in the prompt**: By default, the first sentence is the task description shown in Figure 1. This feature starts the prompt with *You are an excellent <specialist>. You can identify all the mentions of <entity-type> in a sentence, by putting them in a specific format. Here are some examples you can handle:* instead. The *<specialist>* is a *linguist* or a *clinician*, following the task domain.

6. **Include label definitions in the prompt**: This feature adds a one-sentence description for each entity type. Full entity descriptions used can be found in appendix A.3.

7. **Introductory sentence for the test instance**: By default, the demonstrations are immediately followed by the test instance. This feature separates them with *Identify all the mentions of <entity-type> in the following sentence, by putting <begin-tag> in front and a <end-tag> behind each of them.*

8. **Require a long answer for the self-verification**: By default, the self-verification prompt demonstrates *Yes* (respectively *No*) as answers. This feature demonstrates *<mention> is a(n) <entity-type>, yes.* (respectively *<mention> is not a(n) <entity-type>, no.*) instead.

9. **Dialogue template**: This feature replaces the *Input:* and *Output:* in the prompt by dashes to imitate a dialog template.

Identification of optimal prompt configuration In-context learning performance is shown to vary greatly depending on the exact phrasing of prompts (Lu et al., 2022; Min et al., 2022). In addition, the optimal choice for each of these features can vary depending on the model used. For instance, intuitively, models that are heavily pretrained on the English language tend to perform better with an English template than one in the language of the corpus.

While our system aims to search for the best combination of parameters for each model, a grid search over them would require $2^9 = 512$ experiments for each model, for each dataset. In order to build a lighter system, we opt for a greedy search. We iterate over the features and select options that perform better than the default. In Section A.1.4, we illustrate how the optimal parameter combination can be obtained with a greedy search using 5% of the computation required for grid search. Many efforts towards "few-shot learning" with LLMs design prompts on large held-out validation datasets (Brown et al., 2020; Tam et al., 2021; Radford et al., 2021; Qin and Eisner, 2021). This leads to results that are shown (Perez et al., 2021) to be over-optimistic. A true few-shot evaluation involves optimization with access to a small number of annotated instances, which corresponds to our $k = 100$. In this no-training context, we follow (Perez et al., 2021) optimizing these features

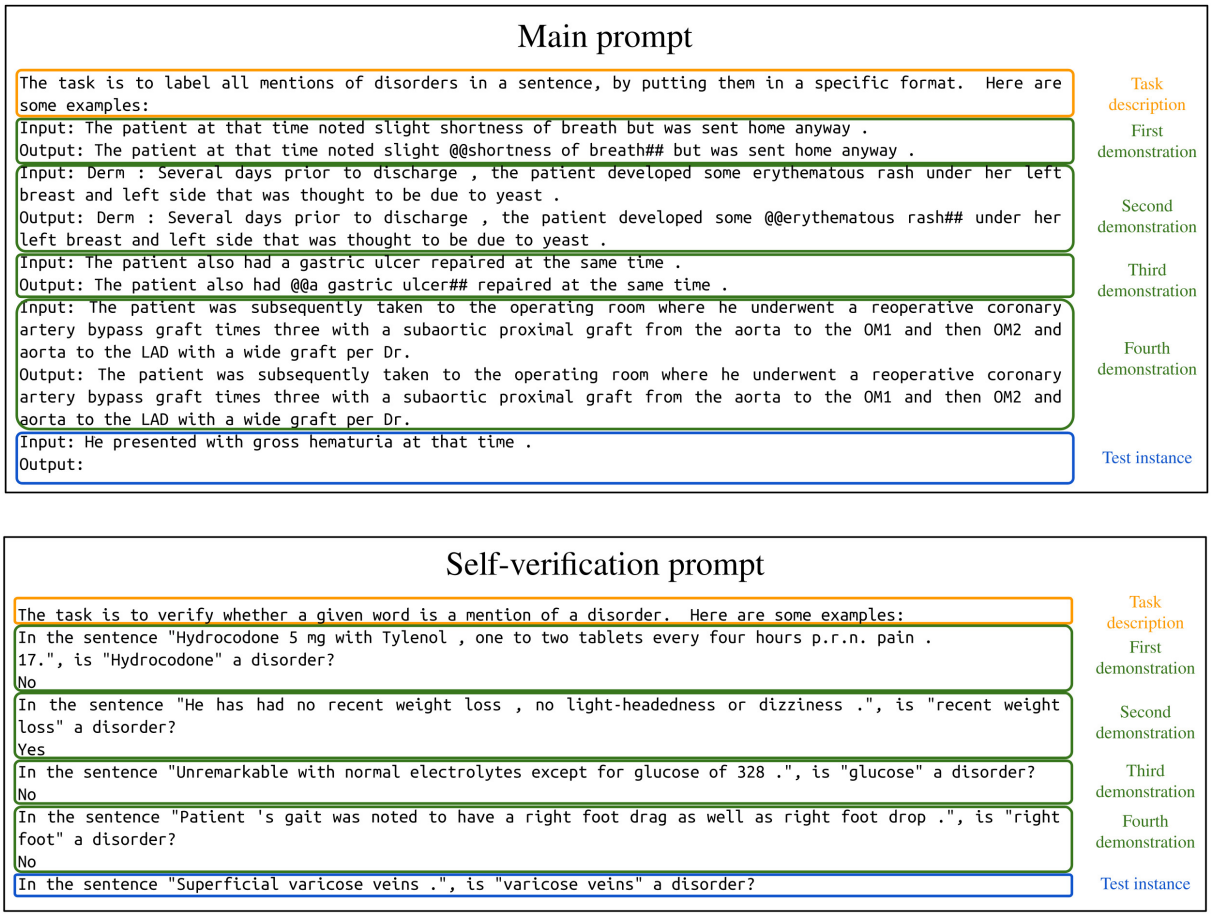

Figure 1: Example of a tagging prompt, used in the main experiment (top) and a self-verification prompt (bottom) for detecting DISO mentions in **n2c2-2019**

through a leave-one-out cross-validation (LOOCV) over the validation set.

3.5 Performance metrics

Micro-F1 For simplicity, we evaluate models over one global information extraction performance score. It is computed as the micro-average of F1-measures for each entity type.

Carbon footprint We use GreenAlgorithms v2.2 (Lannelongue et al., 2021) [3] to estimate the carbon footprint of each experiment, based on factors such as runtime, computing hardware and location where electricity used by our computer facility was produced.

4 Results

4.1 Environmental Impact

Figure 2 compares carbon emission incurred by resolving ConLL-2003 using LLM prompting (using

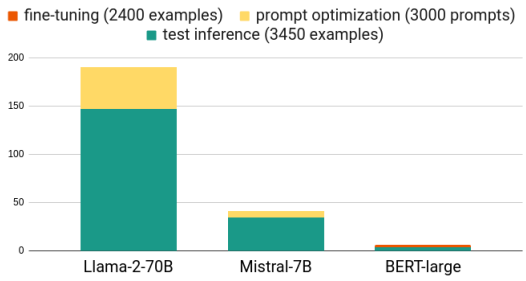

Figure 2: Carbon emission (g) incurred by resolving ConLL-2003 using three models

Llama-2-70B and Mistral-7B) as well as with fine-tuning BERT-large. For comparison, the impact of using Llama-2-70B is comparable to that of driving a thermal car of average size for 1 kilometer.

Appendix A.4 details the carbon emission estimations for all of our experiments. In total, the experiments described in this paper are estimated to have generated around 31kg of CO2 equivalent (29kg for main experiments, and 2kg for ablation).

[3] http://calculator.green-algorithms.org/

	#	Model	English					French					Spanish			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Few-shot approaches																
Causal	1	Llama-2-70B	0.728	0.721	0.312	**0.309**	**0.400**	0.740	0.400	0.483	0.201	0.312	**0.805**	0.616	0.021	0.339
	2	Llama-3-8B-Instruct	**0.756**	**0.727**	**0.478**	0.252	0.390	**0.796**	**0.508**	**0.652**	**0.443**	**0.433**	0.784	**0.764**	**0.226**	**0.437**
	3	Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374
	4	Phi-3-medium-instruct	0.464	0.302	0.234	0.112	0.178	0.394	0.322	0.317	0.139	0.161	0.612	0.461	0.045	0.280
	5	BLOOM-7B1	0.524	0.557	0.279	0.113	0.151	0.148	0.206	0.320	0.197	0.120	0.470	0.419	0.051	0.117
	6	Falcon-40B	0.686	0.708	0.280	0.279	0.305	0.662	0.456	0.378	0.279	0.283	0.720	0.543	0.072	0.267
	7	GPT-J-6B	0.521	0.493	0.167	0.179	0.238	0.423	0.244	0.334	0.080	0.177	0.005	0.142	0.021	0.162
	8	OPT-66B	0.608	0.495	0.227	0.157	0.234	0.624	0.406	0.019	0.206	0.283	0.166	0.273	0.043	0.204
	9	Vicuna-13B	0.657	0.708	0.355	0.236	0.300	0.677	0.350	0.399	0.207	0.326	0.744	0.250	0.040	0.213
	10	Vicuna-7B	0.594	0.489	0.259	0.147	0.172	0.591	0.277	0.439	0.152	0.296	0.659	0.569	0.042	0.151
	11	BioMistral-7B	0.414	0.354	0.175	0.086	0.257	0.547	0.299	0.350	0.236	0.186	0.578	0.540	0.014	0.226
	12	Medalpaca-7B	0.537	0.586	0.272	0.138	0.132	0.529	0.142	0.259	0.162	0.252	0.581	0.490	0.088	0.220
	13	Vigogne-13B	0.593	0.655	0.252	0.176	0.309	0.515	0.250	0.464	0.099	0.142	0.580	0.561	0.010	0.198
Masked	14	mBERT	0.768	0.804	0.624	0.378	0.401	0.801	0.728	0.741	0.588	0.428	**0.812**	0.760	0.324	0.432
	15	XLM-R-large	0.786	0.826	**0.637**	0.462	0.471	0.811	0.781	0.762	0.629	0.531	0.797	0.781	0.325	0.528
	16	BERT-large	0.776	0.814	0.626	0.435	0.422	-	-	-	-	-	-	-	-	-
	17	RoBERTa-large	**0.790**	**0.829**	0.626	0.462	**0.552**	-	-	-	-	-	-	-	-	-
	18	Bio_ClinicalBERT	0.528	0.542	0.621	0.469	0.420	-	-	-	-	-	-	-	-	-
	19	ClinicalBERT	0.462	0.597	0.622	**0.480**	0.397	-	-	-	-	-	-	-	-	-
	20	MedBERT	0.613	0.673	0.607	0.478	0.504	-	-	-	-	-	-	-	-	-
	21	CamemBERT-large	-	-	-	-	-	**0.829**	**0.793**	0.768	**0.661**	**0.564**	-	-	-	-
	22	FlauBERT-large	-	-	-	-	-	0.826	0.778	0.760	0.635	0.540	-	-	-	-
	23	DrBERT-4GB	-	-	-	-	-	0.587	0.599	0.730	0.602	0.497	-	-	-	-
	24	CamemBERT-bio	-	-	-	-	-	0.782	0.761	**0.779**	0.636	0.557	-	-	-	-
	25	BETO	-	-	-	-	-	-	-	-	-	-	0.794	0.732	0.352	0.522
	26	PatanaBERT	-	-	-	-	-	-	-	-	-	-	0.802	0.769	0.343	0.487
	27	TulioBERT	-	-	-	-	-	-	-	-	-	-	0.804	**0.798**	0.340	0.482
	28	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	0.804	0.758	0.354	**0.578**
	29	BSC-Bio	-	-	-	-	-	-	-	-	-	-	0.804	0.775	**0.358**	0.552
Masked fully-supervised (skyline)																
		RoBERTa-large	0.919	0.939	0.718	0.712	0.815	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	0.928	0.834	0.828	0.748	0.713	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	0.918	0.881	0.411	0.736

Table 1: This table presents the micro-F1 obtained from few-shot experiments. Skyline results are obtained using all training data available instead of the few-shot setting.

4.2 Comparison of model performance

Table 1 presents the micro-F1 performance of the models on the test set of each dataset. Figures 3 to 5 present the results for each language with a vizualization of model type (circles for causal models vs. squares for masked models), model size in terms of parameters and domain (clinical in green vs. general in blue). Overall, results show that, despite being smaller and theoretically requiring a larger amount of training data, masked, "BERT-like" models consistently outperform generative LLM prompting in the context of few-shot NER commonly found in biomedical applications. Specifically, on English and Spanish, Figures 3 and 5, show that, while some generative LLMs are competitive in the general domain, they suffer a sharper performance drop in clinical NER, compared to fine-tuned MLMs. On French, Figure 4 suggests that generative LLM prompting does not seem competitive on either domain.

Moreover, this performance comes at a much lower environmental impact (CO2 emissions are 10-50 times lower for MLMs vs. prompted LLMs). Another important finding is that, in addition to high performance scores, MLMs achieve results that are relatively close to each other. For example, on the WikiNER generalist task in English, the 4 general-domain models tested achieved F1-scores of between 0.768 and 0.79.

Besides, we show that domain-adapted MLMs (e.g., ClinicalBERT, CamemBERT-bio) exhibit a sharp performance drop in general domain tasks, illustrating the classical issue of "catastrophic forgetting". In addition, they do not contribute performance improvements in specialized tasks, with the exception of Spanish tasks. However, there is a notable difference in model size: specialized models only have 110 million parameters (vs. 340 million for other models).

Named entity recognition based on BERT-type representations has received a great deal of attention in recent years, and is undoubtedly more mature than the use of LLM prompting for this task. We have implemented the prompt-based NER techniques recently published in the literature, to the best of our knowledge. It is, of course, possible that new approaches will make it possible to increase performance in the future. However, this is arguably a difficult task for a generative model, as it is highly constrained in its syntax and its evaluation.

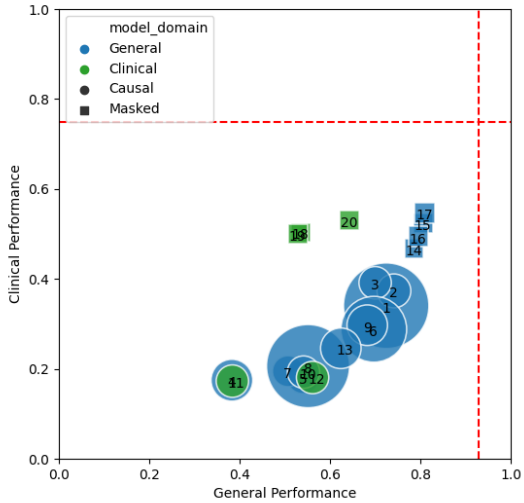

Figure 3: Performance of models on English. The general performance is the average of micro-F1 obtained on WikiNER-en and CoNLL-2003. The clinical performance is the average on E3C-en, n2c2 and NCBI-Disease. The red lines represent the *skyline* performance obtained with the entirety of each training dataset.

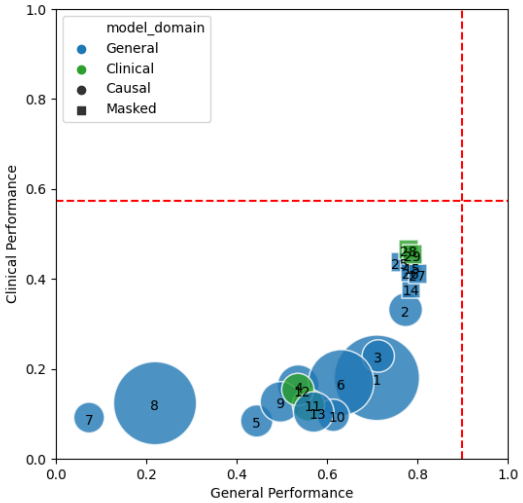

Figure 5: Performance of models on Spanish. The general performance is the average of micro-F1 obtained on WikiNER-es and CoNLL-2002. The clinical performance is the average on E3C-es and CWL. The red lines represent the *skyline* performance obtained with the entirety of each training dataset.

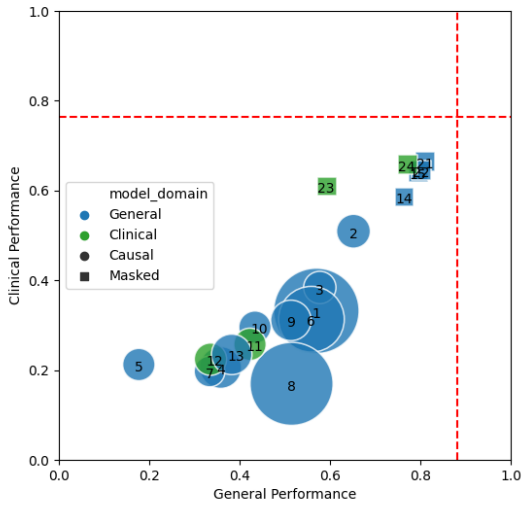

Figure 4: Performance of models on French. The general performance is the average of micro-F1 obtained on WikiNER-fr and QuaeroFrenchPress. The clinical performance is the average on E3C-fr, EMEA and MEDLINE. The red lines represent the *skyline* performance obtained with the entirety of each training dataset.

These results are no indication of performance on other tasks such as classification.

4.3 Practical use of language models for low-resource NER

Overall, our experiments suggest that the performance of language models for clinical named entity recognition is currently sub-optimal. In particular, even MLM-based models, simply fine-tuned on the limited data available, fail to approach the performance of fully supervised models. The three large models trained with the entirety of each training dataset (*skylines* Table 1) systematically outperform the best few-shot results, by 5% to 16% for the general domain, and 8% to 48% for the biomedical domain. However, performance can be judged satisfactory enough for pre-annotation use, to complement or accelerate manual annotation, for example in an online or active learning context.

5 Conclusion

This study assessed the performance of two types of large languages models, for few-shot entity recognition in three languages. Our experiments show that few-shot learning performance is significantly lower in the clinical vs. general domain. While masked language models perform better than causal language models (higher F1, lower CO2 emissions), LLM prompting is not yet suited for effective information extraction.

Limitations of our study

Random Noise and Significance In MLM experiments, the parameters of the NER tagging layer added on top of the pretrained language model are initialized randomly. Likewise, in LLM prompting experiments, the demonstrations in the prompts are shuffled randomly, and the negative examples in the self verification prompts are selected randomly. These random decisions can introduce noise in our performance measurements. Replicating all the experiments would allow us to draw more solid conclusions (Reimers and Gurevych, 2017), but would also come at a considerable cost (29kg of CO2 equivalent, and around 64 hours of computation for each replication). The large number of models tested and tasks addressed can however support the main observations of this article. For instance, we use Almost Stochastic Order (ASO) [4] (Dror et al., 2019) with a confidence level $\alpha = 0.05$ to measure the significance of the superiority of fine-tuned MLMs over prompted LLMs for each dataset separately. We do not always observe satisfying values of ϵ_{min} as to whether MLMs perform better than prompted LLMs on general-domain NER (0.426, 0.081 and 0.028 respectively on WikiNER-English, CoNLL2003 and WikiNER-French). Regarding clinical NER, MLMs perform significantly better than prompted LLMs : MLMs are stochastically dominant over prompted LLMs (ϵ_{min}=0) for all clinical datasets.

Data contamination The size of the training corpora used for creating LLMs makes it increasingly difficult to control for data contamination, i.e. the presence of test corpora. Moreover, (Balloccu et al., 2024) describe the problem of "indirect contamination", encountered when models are iteratively improved by using data coming from users, including test dataset.

The community is calling for efforts towards better documentation of training datasets (Bender and Friedman, 2018). While some datasets are by construction incompatible with some models (e.g., there is no Spanish training corpus in GPT-J or Llama-2) we are unable to affirm full exclusion of all datasets from all models studied.

Fine-tuning generative models Our study focuses on low-data scenarios prevalent in real-world biomedical applications, where computational resources, notably GPU availability, are constrained. Additionally, the use of models via API calls is challenging due to privacy and security concerns. Consequently, we selected methods requiring limited computing power : either fine-tuning "small", BERT-like models, or prompting larger, frozen, models. This pragmatic choice stems from the need to balance performance with computational efficiency, ensuring that the proposed methods remain feasible for implementation in resource-limited settings.

However, more computationally expensive solutions may exist and may be effectively deployed in other environments. Typically, one could use the limited available data to partially fine-tune a generative LLM for NER (Liao et al., 2023).

Error analysis To facilitate comparison between models, we evaluated models over one global information extraction performance score: the micro-F1 measure. Overall, precision and recall scores (not shown) are balanced across entity types. In clinical datasets, some entity types, such as chemicals and anatomy, yield higher performance than others such as devices and procedures. This disparity cannot be explained by the distribution in the training corpus, since prompts provide the same number of examples across entity types. We hypothesize that the regularity and prevalence in LLM pretaining corpora of some entities may explain the higher performance. In future work, we plan more in-depth analysis, to better understand the performance of the LLM prompting approach.

Ethical considerations

All the datasets analyzed in our experiments are publicly available corpora, used consistently with the relevant data use agreements.

The use of LLMs can incur significant environmental impact. We have measured carbon emissions using GreenAlgorithms. Our experiments support the conclusion that MLMs yield higher information extraction performance with lower carbon emission, compared to LLM prompting for NER.

[4] Given the performance scores of two algorithms A and B, each of which run several times with different settings, ASO computes a test-specific value (ϵ_{min}) that indicates how far algorithm A is from being significantly better than algorithm B. If distance ϵ_{min} = 0.0, one can claim that A stochastically dominant over B with the predefined significance level. The literature commonly interprets ϵ_{min} < 0.5 as an indicator of a significant superiority of A over B.

Acknowledgements

This work has received funding from the French "Agence Nationale pour la Recherche" under grant agreement CODEINE ANR-20-CE23-0026-01. This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014533). The authors thank Dr. Juan Manual Coria for his help phrasing prompts in Spanish.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Rami Aly, Andreas Vlachos, and Ryan McDonald. 2021. Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1516–1528, Online. Association for Computational Linguistics.

Dhananjay Ashok and Zachary C. Lipton. 2023. PromptNER: Prompting For Named Entity Recognition.

Samy Ateia and Udo Kruschwitz. 2023. Is ChatGPT a Biomedical Expert? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks.

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*, 52:571–601.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han, and Le Sun. 2023a. Learning In-context Learning for Named Entity Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13661–13675, Toronto, Canada. Association for Computational Linguistics.

Peng Chen, Jian Wang, Hongfei Lin, Di Zhao, and Zhihao Yang. 2023b. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. *Bioinformatics*, 39(8):btad496.

Hyejin Cho, Wonjun Choi, and Hyunju Lee. 2017. A method for named entity normalization in biomedical articles: Application to diseases and plants. *BMC Bioinformatics*, 18.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Zhaojian Cui, Kai Yu, Zhenming Yuan, Xiaofeng Dong, and Weibin Luo. 2024. Language inference-based learning for Low-Resource Chinese clinical named entity recognition using language model. *Journal of Biomedical Informatics*, 149:104559.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. Biomedical Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep Dominance - How to Properly Compare Deep Neural Models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, and Anne-Sophie Jannot. 2017. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC medical informatics and decision making*, 17(1):1–10.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-Shot Classification in Named Entity Recognition Task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 993–1000, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Álvaro García-Barragán, Alberto González Calatayud, Oswaldo Solarte-Pabón, Mariano Provencio, Ernestina Menasalvas, and Víctor Robles. 2024. GPT for medical entity recognition in Spanish. *Multimedia Tools and Applications*, pages 1–20.

Yao Ge, Yuting Guo, Sudeshna Das, Mohammed Ali Al-Garadi, and Abeed Sarker. 2023. Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, 144:104458.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based Example Selection for In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Christel Gérardin, Perceval Wajsbürt, Pascal Vaillant, Ali Bellamine, Fabrice Carrat, and Xavier Tannier. 2022. Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, 128:102311.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. MedAlpaca–An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247*.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online. Association for Computational Linguistics.

Niu Hu, Xuan Zhou, Bing Xu, Hanqing Liu, Xiangjin Xie, and Hai-Tao Zheng. 2023a. VPN: Variation on Prompt Tuning for Named-Entity Recognition. *Applied Sciences*, 13(14).

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023b. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021a. Few-Shot Named Entity Recognition: An Empirical Baseline Study. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021b. Few-Shot Named Entity Recognition: An Empirical Baseline Study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Amirhossein Layegh, Amir H Payberah, Ahmet Soylu, Dumitru Roman, and Mihhail Matskin. 2023. ContrastNER: Contrastive-based Prompt Tuning for Few-shot NER. *arXiv preprint arXiv:2305.17951*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2700, Dublin, Ireland. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. on Knowl. and Data Eng.*, 34(1):50–70.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Baohao Liao, Yan Meng, and Christof Monz. 2023. Parameter-Efficient Fine-Tuning without Introducing New Latency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4260, Toronto, Canada. Association for Computational Linguistics.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Andy T. Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew O. Arnold. 2022. QaNER: Prompting Question Answering Models for Few-shot Named Entity Recognition. *ArXiv*, abs/2203.01543.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yang Liu. 2019. Fine-tune BERT for Extractive Summarization.

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

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. MCN: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103132.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022a. Label Semantics for Few Shot Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. 2022b. Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732, Seattle, United States. Association for Computational Linguistics.

Bernardo Magnini, Begona Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2021. The E3C Project: European Clinical Case Corpus. *Language*, 1(L2):L3.

Antoine Magron, Anna Dai, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. 2024. JobSkape: A framework for generating synthetic job postings to enhance skill matching. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 43–58, St. Julian's, Malta. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Alexa McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. *Studies in health technology and informatics*, 84:216–20.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, pages 24–30.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models. In *Advances in Neural Information Processing Systems*.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How Context Affects Language Models' Factual Predictions.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Stefan Schweter and Alan Akbik. 2021. FLERT: Document-Level Features for Named Entity Recognition.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and Simplifying Pattern Exploiting Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv preprint arXiv:2305.12031*.

Rian Touchent and Éric de la Clergerie. 2024. CamemBERT-bio: Leveraging continual pre-training for cost-effective models on French biomedical data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2692–2701, Torino, Italia. ELRA and ICCL.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Charangan Vasantharajan, Kyaw Zin Tun, Ho Thi-Nga, Sparsh Jain, Tong Rong, and Chng Eng Siong. 2022. MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1482–1488.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Perceval Wajsbürt. 2021. *Extraction and normalization of simple and structured entities in medical documents*. Theses, Sorbonne Université.

Perceval Wajsbürt, Arnaud Sarfati, and Xavier Tannier. 2021. Medical concept normalization in French using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, 114:103684.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023a. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, pages 1–10.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. GPT-NER: Named Entity Recognition via Large Language Models.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. 2024. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, 5(1):100887.

Yi Yang and Arzoo Katiyar. 2020. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

Feiyang Ye, Liang Huang, Senjie Liang, and KaiKai Chi. 2023. Decomposed Two-Stage Prompt Learning for Few-Shot Named Entity Recognition. *Information*, 14(5).

Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. 2024. Prompt engineering paradigms for medical applications: Scoping review. *J Med Internet Res*, 26:e60501.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendix

A.1 Ablation

To better understand the contribution of each step of our approach, we carried out a series of complementary experiments.

A.1.1 Sample and sample size

We tested our approach with different samples and different sample sizes for one MLM : XLM-RoBERTa-large, and one prompted LLM : Mistral-7B. The results are reported in table 2. It can be noted that, whereas the standard deviation with respect to p is rather high, a significant difference can still be consistently observed between the two models across samples of the same size. We also observe that, as the number of annotated instances decreases, the performance of the MLM drops faster than that of the prompted LLM.

	CoNLL2003			n2c2		
100 annotated instances						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.646	0.626	0.714	0.291	0.178	0.215
XLM-R-large	0.826	0.814	0.786	0.462	0.478	0.526
50 annotated instances						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.615	0.648	0.637	0.278	0.176	0.106
XLM-R-large	0.697	0.77	0.714	0.431	0.476	0.35
25 annotated instances						
	p=1	*p=2*	*p=3*	*p=1*	*p=2*	*p=3*
Mistral-7B	0.509	0.599	0.52	0.152	0.252	0.116
XLM-R-large	0.487	0.588	0.637	0.393	0.361	0.283

Table 2: F1 scores obtained over experiments with different training samples and different training sample sizes.

A.1.2 Listing prompts

In this section, we compare the adopted tagging prompts to listing prompts. In listing prompts, demonstrations simply list the tagged mentions. The list separator is optimized (in the same way as the taggers) between a comma and a newline character. Eventually, the introductory sentences asks to list entities. The results shown in table 3 further corroborate our choice of only focusing on tagging prompts.

A.1.3 Number of demonstrations

The decision to limit the number of annotated examples presented to generative models in the prompt to a maximum of 10 is dictated by two constraints.

Firstly, models impose a limit on the number of input tokens over which attention is calculated. Most models used have a limit of 2048 tokens, but more permissive models, such as Mistral-7B, allow up to 8096 tokens. This constraint translates into a limit on the number of sentences that can be presented in the prompt, ranging from 40 to 50 for Mistral-7B and from 10 to 15 for less permissive models, depending on the corpora and tokenizers. For instance, consider the task of detecting body parts in the French section of E3C. Using Mistral-7B (and its tokenizer), the practical limit is around 40 examples, resulting in an average prompt of 7779.5 tokens. Using Vicuna-13B (and its tokenizer), the limit is around 11 examples, resulting in an average of 1851.5 tokens in the prompt.

Secondly, the improvement brought by adding more examples does not appear to be significant, as observed in Table 4, which shows the results obtained with Mistral-7B when the number of annotated examples is tripled. It is noteworthy that the marginal improvements achieved by tripling the number of examples come at a considerable cost, especially given the quadratic complexity with respect to the length of the prompt.

Therefore, we choose to limit the prompt to 5-10 examples, selected based on the chosen criteria (TF-IDF proximity to the reference sentence or the number of entities present). Instead of selecting examples independently, Gupta et al. (2023) propose selecting them interdependently to improve the representativeness of the prompt. This approach would be interesting to implement in our system in future work.

A.1.4 Hyperparameter grid search

In order to assess the quality of our adopted search method used to find the best feature combination to incorporate in the prompt, we compare this method to a naïve grid search over these features. We test all 512 combinations of our identified 9 features, for Mistral-7B over ConLL2003. The scores found through LOOCV vary between 0.0 and 0.656 with a mean value of 0.387 and a median of 0.46. The best-preforming combination is : *Additional sentences*, *Self-verification*, *Introductory sentence for the test instance* and *Require a long answer for the self-verification*, which is exactly the same combination we found initially through a greedy, tree search, that is around 20 times faster and less consuming.

	English					French					Spanish			
Model	WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Listing prompts														
Mistral-7B	0.659	0.533	0.417	0.281	0.340	0.676	0.083	0.451	0.169	0.403	0.697	0.620	0.211	0.273
Tagging prompts														
Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374

Table 3: F1 scores obtained with the listing and tagging prompts.

	English					French					Spanish			
Model	WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
5/10 demonstrations														
Mistral-7B	0.754	0.646	0.488	0.291	0.395	0.727	0.428	0.590	0.229	0.333	0.720	0.707	0.083	0.374
15/30 demonstrations														
Mistral-7B	0.763	0.692	0.453	0.263	0.377	0.782	0.355	0.587	0.237	0.396	0.785	0.751	0.163	0.413

Table 4: F1 scores obtained with 5/10 demonstations vs. with 15/30 demonstations

A.2 Evaluated models

Table 5 specifies relevant information about the tested models. In the training corpus column, we point out the data contaminations we know of. For instance, MedBERT's documentation explicitly mentions N2C2 as part of its training data. This might lead to artificially high evaluation metrics, as the model is not generalizing to unseen data, but instead leveraging pre-learned information, which compromises the validity of the results in real-world applications.

A.3 NER labels descriptions

Tables 6 to 10 definitions of all the labels present in the studied datasets. These definitions were drawn from available annotation schemes and further curated by the authors and native speakers. These definitions were included in the prompt when the "Include label definitions in the prompt" feature was activated.

A.4 Carbon footprint

Tables 11 and 12 detail the carbon emission estimations for all of our experiments. These estimations were computed with GreenAlgorithms v2.2 (Lannelongue et al., 2021) [5], based on factors such as runtime, computing hardware and location where electricity used by our computer facility was produced. In total, the experiments described in this paper are estimated to have generated around 31kg of CO2 equivalent (29kg for the main experiments, and 2kg for ablation).

[5] http://calculator.green-algorithms.org/

	#	Model	Number of parameters	Training data size	Training corpus
Causal	1	Llama-2-70B[en] (Touvron et al., 2023)	70B	2 trillion tokens	A mix of publicly available online data, mainly in English
	2	Llama-3-8B-Instruct[en]	8B	over 15 trillion tokens	"A new mix" of publicly available online data, mainly in English
	3	Mistral-7B[?] (Jiang et al., 2023)	7B	Undisclosed	Undisclosed
	4	Phi-3-medium-instruct[en] (Abdin et al., 2024)	14B	4.8 trillion tokens	A combination of publicly available corpora, synthetic data and chat format supervised data, mainly in English
	5	BLOOM-7B1[en][fr][es] (Workshop et al., 2022)	7B	1.6 TB	ROOTS (Laurençon et al., 2022), a mix of datasets and pseudo-crawled data 59 languages
	6	Falcon-40B[en][fr][es]	40B	1 trillion tokens	RefinedWeb (Penedo et al., 2023), a dataset of filtered and deduplicated web data
	7	GPT-J-6B[en] (Wang and Komatsuzaki, 2021)	6B	825 GiB	The Pile (Gao et al., 2020), a mixture of public datasets and web data in English
	8	OPT-66B[en] (Zhang et al., 2022)	66B	180 billion tokens	Crawled data from the web, mainly in English
	9	Vicuna-13B[en]* (Zheng et al., 2023)	13B	125K conversations	Llama 2, fine-tuned on conversations collected from ShareGPT.com, mainly in English
	10	Vicuna-7B[en]* (Zheng et al., 2023)	7B	125K conversations	Llama 2, fine-tuned on conversations collected from ShareGPT.com, mainly in English
	11	BioMistral-7B[en]* (Labrak et al., 2024)	7B	3 billion tokens	Mistral, fine-tuned on the PMC Open Access Subset
	12	Medalpaca-7B[en]* (Han et al., 2023)	7B	400K Q.A. pairs	Llama 2, fine-tuned on semi-generated medical question-answer pairs in English
	13	Vigogne-13B[fr][en]*	13B	52K instructions	Llama 2, fine-tuned on English instructions automatically translated to French
Masked	14	mBERT[en][fr][es] (Devlin et al., 2019)	110M	Undisclosed	A corpus featuring 104 languages built from undisclosed sources
	15	XLM-R-large[en][fr][es] (Conneau et al., 2020)	355M	2.5 TB	Filtered CommonCrawl data containing 100 languages
	16	BERT-large[en] (Devlin et al., 2019)	345M	3,3 billion words	BookCorpus (Zhu et al., 2015), a dataset consisting of unpublished books and English Wikipedia.
	17	RoBERTa-large[en] (Liu et al., 2019)	355M	160 GiB	BooksCorpus (Zhu et al., 2015), English Wikipedia, and crawled web data
	18	Bio_ClinicalBERT[en] (Alsentzer et al., 2019)	110M	2 million clinical notes	MIMIC-III (Johnson et al., 2016), a database containing electronic health records from hospitalized ICU patients
	19	ClinicalBERT[en] (Wang et al., 2023a)	110M	1.2 billion words	A large multi-center dataset with a corpus built from undisclosed sources
	20	MedBERT[en] (Vasantharajan et al., 2022)	110M	57 million words	Community datasets (including N2C2 (Luo et al., 2020)) and Crawled medical-related articles from Wikipedia
	21	CamemBERT-large[fr] (Martin et al., 2020)	335M	64 billion tokens	OSCAR (Suárez et al., 2020), a corpus of web data in French
	22	FlauBERT-large[fr] (Le et al., 2020)	335M	13 billion tokens	A mix of French Wikipedia, French books, and French web data
	23	DrBERT-4GB[fr] (Labrak et al., 2023)	110M	1 billion words	A mix of publicly available biomedical corpora in French (including QuaeroFrenchMed (Névéol et al., 2014)).
	24	CamemBERT-bio[fr] (Touchent and de la Clergerie, 2024)	110M	413 million words	A mix of publicly available biomedical corpora in French (including E3C (Magnini et al., 2021)).
	25	BETO[es] (Cañete et al., 2020)	110M	3 billion words	Spanish Wikipedia and Spanish data from OPUS (Tiedemann, 2012)
	26	PatanaBERT[es]	110M	Undisclosed	Spanish
	27	TulioBERT[es]	110M	Undisclosed	Spanish
	28	BSC-BioEHR[es] (Carrino et al., 2022)	110M	1.1 billion tokens	A mixture of biomedical community datasets including EHR documents and crawled data in Spanish
	29	BSC-Bio[es] (Carrino et al., 2022)	110M	963 million tokens	A mixture of biomedical community datasets and crawled data in Spanish

Table 5: Characterization of the language models used in our experiments in terms of parameters and training corpus. Models marked with [en] (respectively [fr], [es]) are heavily trained on English (respectively French, Spanish). CLMs marked with * are fine-tuned versions of other CLMs.

Tag	Tag name (in singular)	Description
PER	person names (a person's name)	These are names of persons such as real people or fictional characters.
FAC	facilities (a facility)	These are names of man-made structures such as infrastructure, buildings and monuments.
LOC	locations (a location)	These are names of geographical locations such as landmarks, cities, countries and regions.
ORG	organizations (an organization)	These are names of organizations such as companies, agencies and political parties.
FUNC	functions and jobs (a function or a job)	These are words that refer to a profession or a job.
ACTI	activities and behaviors (an activity or behavior)	These are words that refer to human activities, behaviors or events as well as governmental or regulatory activities.
ANAT	anatomy (an anatomy)	These are words that refer to the structure of the human body, its organs and their position, such as body parts or organs, systems, tissues, cells, body substances and embryonic structures.
CHEM	chemicals and drugs (a chemical or a drug)	These are words that refer to a substance or composition that has a chemical characteristic, especially a curative or preventive property with regard to human or animal diseases, such as drugs, antibiotics, proteins, hormones, enzymes and hazardous or poisonous substances.
CONC	concepts and ideas (a concept or an idea)	These are words that refer to a concept or an idea, such as a classification, an intellectual product, a language, a law or a regulation.
DEVI	medical devices (a device)	These are words that refer to a medical device used to administer care or perform medical research.
DISO	disorders (a disorder)	These are words that refer to an alteration of morphology, function or health of a living organism, animal or plant, such as congenital abnormalities, dysfunction, injuries, signs or symptoms or observations.
GENE	genes and molecular sequences (a gene or a molecular sequence)	These are words that refer to a gene, a genome or a molecular sequence.
GEOG	geographical areas (a geographical area)	These are words that refer to a country, a region or a city.
LIVB	living beings (a living being)	These are words that refer to a living being or a group of living beings, such as a person or a group of persons, a plant or a category of plants, an animal or a category of animals.
OBJC	objects (an object)	These are words that refer to anything animate or inanimate that affects the senses, such as physical manufactured objects.
OCCU	occupations (an occupation)	These are words that refer to a professional occupation or discipline.
ORGA	organizations (an organization)	These are words that refer to an organization such as healthcare related organizations.
PHEN	phenomema (a phenomemon)	These are words that refer to a phenomenon that occurs naturally or as a result of an activity, such as a biologic function.
PHYS	physiology (a physiology)	These are words that refer to any element that contributes to the mechanical, physical and biochemical functioning or organization of living organisms and their components.
PROC	procedures (a procedure)	These are words that refer to an activity or a procedure that contributes to the diagnosis or treatment of patients, the information of patients, the training of medical personnel or biomedical research.
EVENT	events (an event)	These are words that refer to actions, states, and circumstances that are relevant to the clinical history of a patient such as pathologies and symptoms, or more generally words like "enters", "reports" or "continue".

Table 6: Description of the NER tags used in our experiments for English.

Tag	Tag name (in singular)	Description
TIMEX3	time expressions (a time expression)	These are time expressions such as dates, times, durations, frequencies, or intervals.
RML	results and measurements (a result or a measurement)	These are test results, results of laboratory analyses, formulaic measurements, and measure values.
ACTOR	actors (an actor)	These are words that refer patients, healthcare professionals, or other actors relevant to the clinical history of a patient.
Abbreviation	abbreviations (an abbreviation)	These are words that refer to abbreviations.
Body_Part	body parts (a body part)	These are words that refer to organs and anatomical parts of persons.
Clinical_Finding	clinical findings (a clinical finding)	These are words that refer to observations, judgments or evaluations made about patients.
Diagnostic_Procedure	diagnostic procedures (a diagnostic procedure)	These are words that refer to tests that allow determining the condition of the individual.
Disease	diseases (a disease)	These are words that describe an alteration of the physiological state in one or several parts of the body, due to generally known causes, manifested by characteristic symptoms and signs, and whose evolution is more or less predictable.
Family_Member	family members (a family member)	These are words that refer to family members.
Laboratory_or _Test_Result	laboratory or test results (a laboratory or test result)	These are words that refer to any measurement or evaluation obtained from a diagnostic support examination.
Laboratory_Procedure	laboratory procedures (a laboratory procedure)	These are words that refer to tests that are performed on various patient samples that allow diagnosing diseases by detecting biomarkers and other parameters. Blood, urine, and other fluids and tissues that use biochemical, microbiological and/or cytological methods are considered.
Medication	medications (a medication)	These are words that refer to medications or drugs used in the treatment and/or prevention of diseases, including brand names and generics, as well as names for groups of medications.
Procedure	procedures (a procedure)	These are words that refer to activities derived from the care and care of patients.
Sign_or_Symptom	signs or symptoms (a sign or symptom)	These are words that refer to manifestations of a disease, determined by medical examination or perceived and expressed by the patient.
Therapeutic_Procedure	therapeutic procedures (a therapeutic procedure)	These are words that refer to activities or treatments that are used to prevent, repair, eliminate or cure the individual's disease.
CompositeMention	composite mentions of diseases (a composite mention of diseases)	These are words that refer to mentions of multiple diseases, such as "colorectal, endometrial, and ovarian cancers".
DiseaseClass	disease classes (a disease class)	These are words that refer to classes of diseases, such as "an autosomal recessive disease".
Modifier	modifiers (a modifier of diseases)	These are words that refer to modifiers of diseases, such as "primary" or "C7-deficient".
SpecificDisease	diseases (a disease)	These are words that refer to specific diseases, such as "diastrophic dysplasia".

Table 7: Description of the NER tags used in our experiments for English, continued.

Tag	Tag name (in singular)	Description
PER	de noms de personnes (un nom de personne)	Il s'agit des noms de personnes, qu'elles soient réelles ou fictives.
FAC	de productions humaines (une production humaine)	Il s'agit des noms de structures faites par les humains comme des infrastructures, des bâtiments ou des monuments.
LOC	de lieux (un lieu)	Il s'agit des noms de lieux comme des endroits, villes, pays ou régions.
ORG	d'organisations (une organisation)	Il s'agit des noms d'organisations comme des entreprises, des agences ou des partis politiques.
FUNC	de fonctions et métiers (une fonction ou un métier)	Il s'agit de mots qui se rapportent à une activité professionnelle.
ANAT	d'anatomie (une partie du corps)	Il s'agit d'une entité se rapportant à la structure du corps humain, ses organes et leur position. Il s'agit principalement des parties du corpus ou organes, des appareils, des tissus, des cellules, des substances corporelles et des organismes embryonaires.
CHEM	de médicaments et substances chimiques (un médicament ou une substance chimique)	Il s'agit d'une substance ou composition présentant des propriétés chimiques caractéristiques, en particulier des propriétés curatives ou préventives à l'égard des maladies humaines ou animales. Il s'agit principalement des médicaments disponibles en pharmacie, des antibiotiques, des proteines, des hormones, des substances dangereuses, des enzymes.
DEVI	de matériel (un matériel)	Il s'agit d'un matériel utilisé pour administrer des soins ou effectuer des recherches médicales.
DISO	de problèmes médicaux (un problème médical)	Il s'agit d'une altération de la morphologie, des fonctions, ou de la santé d'un organisme vivant, animal ou végétal. Il peut s'agir de malformations, de maladies, de blessure, de signe ou symptome ou d'une observation.
GEOG	de zones géographiques (une zone géographique)	Il s'agit d'un pays, une région, ou une ville.
LIVB	d'êtres vivants (un être vivant)	Il s'agit d'un être vivant ou groupe d'êtres vivants. Il peut s'agir d'une personne ou d'un groupe de personnes, d'une plante ou d'une catégorie de végétaux, d'un animal ou d'une catégorie d'animaux.
OBJC	d'objets (un objet)	Il s'agit de tout ce qui, animé ou inanimé, affecte les sens. Ici, il s'agit principalement d'objets physiques manufacturés.
PHEN	de phénomènes (un phénomène)	Il s'agit d'un phénomène qui se produit naturellement ou à la suite d'une activité. Il s'agit principalement de fonctions biologiques.
PHYS	de physiologie (une physiologie)	Il s'agit de tout élément contribuant au fonctionnement ou à l'organisation mécanique, physique et biochimique des organismes vivants et de leurs composants.
PROC	de procédures (une procédure)	Il s'agit d'une activité ou procédure contribuant au diagnostic ou au traitement des patients, à l'information des patients, la formation du personnel médical ou à la recherche biomédicale.
EVENT	d'événements (un événement)	Il s'agit d'une action, d'un état ou d'une circonstance qui est pertinent pour l'histoire clinique d'un patient. Il peut s'agir de pathologies et symptômes, ou plus généralement de mots comme "entre", "rapporte" ou "continue".
TIMEX3	d'expressions temporelles (une expression temporelle)	Il s'agit d'expressions temporelles comme des dates, heures, durées, fréquences, ou intervalles.
RML	de résultats et mesures (un résultat ou une mesure)	Il s'agit de résultats d'analyses de laboratoire, de mesures formelles, et de valeurs de mesure.
ACTOR	d'acteurs (un acteur)	Il s'agit de patients, de professionnels de santé, ou d'autres acteurs pertinents pour l'histoire clinique d'un patient.

Table 8: Description of the NER tags used in our experiments for French.

Tag	Tag name (in singular)	Description
PER	nombres de personas (un nombre de persona)	Estos son nombres de personas, ya sean reales o personajes ficticios.
FAC	instalaciones (una instalación)	Estos son nombres de estructuras hechas por el hombre como infraestructura, edificios y monumentos.
LOC	lugares (un lugar)	Estos son nombres de ubicaciones geográficas como hitos, ciudades, países y regiones.
ORG	organizaciones (una organización)	Estos son nombres de organizaciones como empresas, agencias y partidos políticos.
ACTI	actividades y comportamientos (una actividad o comportamiento)	Estas son palabras que se refieren a actividades humanas, comportamientos o eventos, así como actividades gubernamentales o regulatorias.
ANAT	anatomía (una anatomía)	Estas son palabras que se refieren a la estructura del cuerpo humano, sus órganos y su posición, como partes del cuerpo u órganos, sistemas, tejidos, células, sustancias corporales y estructuras embrionarias.
CHEM	productos químicos y medicamentos (un producto químico o un medicamento)	Estas son palabras que se refieren a una sustancia o composición que tiene una característica química, especialmente una propiedad curativa o preventiva con respecto a las enfermedades humanas o animales, como medicamentos, antibióticos, proteínas, hormonas, enzimas y sustancias peligrosas o venenosas.
CONC	conceptos e ideas (un concepto o una idea)	Estas son palabras que se refieren a un concepto o una idea, como una clasificación, un producto intelectual, un idioma, una ley o un reglamento.
DEVI	dispositivos médicos (un dispositivo)	Estas son palabras que se refieren a un dispositivo médico utilizado para administrar atención o realizar investigaciones médicas.
DISO	trastornos (un trastorno)	Estas son palabras que se refieren a una alteración de la morfología, la función o la salud de un organismo vivo, animal o vegetal, como anomalías congénitas, disfunción, lesiones, signos o síntomas u observaciones.
GENE	genes y secuencias moleculares (un gen o una secuencia molecular)	Estas son palabras que se refieren a un gen, un genoma o una secuencia molecular.
GEOG	áreas geográficas (un área geográfica)	Estas son palabras que se refieren a un país, una región o una ciudad.
LIVB	seres vivos (un ser vivo)	Estas son palabras que se refieren a un ser vivo o un grupo de seres vivos, como una persona o un grupo de personas, una planta o una categoría de plantas, un animal o una categoría de animales.
OBJC	objetos (un objeto)	Estas son palabras que se refieren a cualquier cosa animada o inanimada que afecte los sentidos, como objetos físicos fabricados.
OCCU	ocupaciones (una ocupación)	Estas son palabras que se refieren a una ocupación o disciplina profesional.
ORGA	organizaciones (una organización)	Estas son palabras que se refieren a una organización, por ejemplo organizaciones relacionadas con la salud.
PHEN	fenómenos (un fenómeno)	Estas son palabras que se refieren a un fenómeno que ocurre naturalmente o como resultado de una actividad, por ejemplo una función biológica.

Table 9: Description of the NER tags used in our experiments for Spanish.

Tag	Tag name (in singular)	Description
PHYS	fisiología (una fisiología)	Estas son palabras que se refieren a cualquier elemento que contribuya al funcionamiento mecánico, físico y bioquímico o la organización de los organismos vivos y sus componentes.
PROC	procedimientos (un procedimiento)	Estas son palabras que se refieren a una actividad o un procedimiento que contribuye al diagnóstico o tratamiento de pacientes, la información de pacientes, la capacitación del personal médico o la investigación biomédica.
EVENT	eventos (un evento)	Estas son palabras que se refieren a acciones, estados y circunstancias que son relevantes para la historia clínica de un paciente, como patologías y síntomas, o más generalmente palabras como "entra", "reporta" o "continúa".
TIMEX3	expresiones de tiempo (una expresión de tiempo)	Estas son expresiones de tiempo como fechas, horas, duraciones, frecuencias o intervalos.
RML	resultados y mediciones (un resultado o una medida)	Estos son resultados de análisis de laboratorio, mediciones formales y valores de medición.
ACTOR	actores (un actor)	Estas son palabras que se refieren a pacientes, profesionales de la salud u otros actores relevantes para la historia clínica de un paciente.
Abbreviation	abreviaciones (una abreviación)	Estas son los casos de siglas y acrónimos.
Body_Part	partes del cuerpo (una parte del cuerpo)	Estas son palabras que se refieren a òrganos y partes anatómicas de personas.
Clinical_Finding	hallazgos clínicos (un hallazgo clínico)	Estas son palabras que se refieren a observaciones, juicios o evaluaciones que se hacen sobre los pacientes.
Diagnostic_Procedure	procedimientos diagnósticos (un procedimiento diagnóstico)	Estas son palabras que se refieren a exámenes que permiten determinar la condición del individuo.
Disease	enfermedades (una enfermedad)	Estas son palabras que describen una alteración del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, y cuya evolución es más o menos previsible.
Family_Member	miembros de la familia (un miembro de la familia)	Estas son palabras que se refieren a miembros de la familia.
Laboratory_or _Test_Result	resultados de exámenes de laboratorio u otras pruebas (un resultado de un examen de laboratorio u otra prueba)	Estas son palabras que se refieren a cualquier medición o evaluación obtenida a partir de un exámen de apoyo diagnóstico.
Laboratory_Procedure	procedimientos de laboratorio (un procedimiento de laboratorio)	Estas son palabras que se refieren a exámenes que se realizan en diversas muestras de pacientes que permiten diagnosticar enfermedades mediante la detección de biomarcadores y otros parámetros. Se consideran los análisis de sangre, orina, y otros fluidos y tejidos que emplean métodos bioquímicos, microbiológicos y/o citológicos.
Medication	medicamentos o drogas (un medicamento o una droga)	Estas son palabras que se refieren a medicamentos o drogas empleados en el tratamiento y/o prevención de enfermedades, incluyendo marcas comerciales y genéricos, así como también nombres para grupos de medicamentos.
Procedure	procedimientos (un procedimiento)	Estas son palabras que se refieren a actividades derivadas de la atención y el cuidado de los pacientes.
Sign_or_Symptom	signos o síntomas (un signo o un síntoma)	Estas son palabras que se refieren a manifestaciones de una enfermedad, determinadas mediante la exploración médica o percibidas y expresadas por el paciente.
Therapeutic_Procedure	procedimientos terapéuticos (un procedimiento terapéutico)	Estas son palabras que se refieren a actividades o tratamientos que es empleado para prevenir, reparar, eliminar o curar la enfrmedad del individuo.

Table 10: Description of the NER tags used in our experiments for Spanish, continued.

	#	Model	English					French					Spanish			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Few-shot approaches																
Causal	1	LLAMA-2-70B	46	44	126	233	54	85	131	129	273	284	41	76	114	344
	2	LLAMA-3-8B-Instruct	3	5	12	19	12	5	7	7	11	23	4	7	12	36
	3	Mistral-7B	4	6	12	24	8	5	8	14	13	25	7	5	11	27
	4	Phi-3-medium-instruct	5	8	14	24	6	12	9	14	17	25	9	15	19	28
	5	BLOOM-7B1	4	6	10	26	9	8	13	9	26	20	4	8	8	18
	6	Falcon-40B	49	45	56	176	45	31	58	75	162	129	33	25	82	99
	7	GPT-J-6B	7	6	8	23	7	5	8	13	21	17	6	6	13	28
	8	OPT-66B	73	50	120	253	96	38	64	138	273	240	57	52	106	247
	9	Vicuna-13B	10	11	20	52	11	11	12	18	33	40	10	11	22	51
	10	Vicuna-7B	6	8	14	17	6	5	10	10	24	14	8	6	13	27
	11	BioMistral-7B	7	8	11	17	5	7	12	10	28	14	8	8	11	23
	12	Medalpaca-7B	8	4	17	24	10	7	14	11	19	13	7	8	15	26
	13	Vigogne-13B	14	14	29	37	11	13	20	26	36	39	11	14	32	44
Masked	11	mBERT	2	1	2	2	2	2	2	2	1	1	1	2	1	2
	12	XLM-R-large	2	2	2	1	2	2	2	2	2	2	1	1	1	2
	13	BERT-large	2	1	2	2	2	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	1	2	2	2	2	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	2	2	1	2	1	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	1	1	2	2	1	-	-	-	-	-	-	-	-	-
	17	MedBERT	2	2	1	1	1	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	1	1	1	2	2	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	2	2	2	2	2	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	2	2	2	2	2	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	1	2	2	2	2	-	-	-	-
	23	BETO	-	-	-	-	-	-	-	-	-	-	2	1	1	1
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	2	2	2	2
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	1	2	2	1
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	2	2	2	2
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	2	2	2	2
Masked fully-supervised (skyline)																
		RoBERTa-large	647	68	5	12	24	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	595	15	4	5	8	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	579	41	3	21

Table 11: This table presents the carbon emissions (in g) of the optimization on the validation set of each model over each dataset. For CLMs, this correponds to the tree search over the prompt features through cross-validation. For MLMs, this corresponds to the supervised fine-tuning and training of the model.

	#	Model	English					French					Spanish			
			WikiNER	CoNLL2003	E3C	n2c2	NCBI	WikiNER	QFP	E3C	EMEA	MEDLINE	WikiNER	CoNLL2002	E3C	CWL
Few-shot approaches																
Causal	1	LLAMA-2-70B	812	147	36	196	33	508	11	13	92	47	514	201	11	198
	2	LLAMA-3-8B-Instruct	240	39	9	63	22	158	3	4	29	21	279	53	2	34
	3	Mistral-7B	234	35	8	59	21	148	3	4	27	20	261	50	2	32
	4	Phi-3-medium-instruct	326	48	12	64	25	380	5	9	62	48	529	70	4	76
	5	BLOOM-7B1	220	33	8	44	16	255	3	5	38	29	261	47	2	46
	6	Falcon-40B	600	109	26	144	46	722	9	19	155	70	752	154	9	157
	7	GPT-J-6B	146	17	4	53	20	245	2	6	14	26	154	40	3	53
	8	OPT-66B	765	139	33	185	63	971	12	27	179	93	993	196	12	217
	9	Vicuna-13B	314	47	11	63	24	363	5	8	61	46	502	67	4	74
	10	Vicuna-7B	146	17	4	53	20	246	2	6	14	26	155	65	3	53
	11	BioMistral-7B	235	35	9	49	17	269	3	5	43	32	272	49	2	48
	12	Medalpaca-7B	192	24	5	39	14	98	2	2	17	13	172	53	1	21
	13	Vigogne-13B	322	49	11	65	24	245	5	6	44	33	361	68	3	66
Masked	11	mBERT	14	4	<1	2	<1	15	1	<1	1	1	13	2	<1	2
	12	XLM-R-large	14	4	<1	2	<1	15	1	<1	1	1	13	2	<1	2
	13	BERT-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	14	RoBERTa-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	15	Bio_ClinicalBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	16	ClinicalBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	17	MedBERT	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
	18	CamemBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	19	FlauBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	20	DrBERT-4GB	-	-	-	-	-	17	1	<1	1	1	-	-	-	-
	21	CamemBERT-bio	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
	22	BETO	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	23	PatanaBERT	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	24	TulioBERT	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	25	BSC-BioEHR	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
	26	BSC-Bio	-	-	-	-	-	-	-	-	-	-	13	2	<1	2
Masked fully-supervised (skyline)																
		RoBERTa-large	14	4	<1	2	<1	-	-	-	-	-	-	-	-	-
		CamemBERT-large	-	-	-	-	-	15	1	<1	1	1	-	-	-	-
		BETO	-	-	-	-	-	-	-	-	-	-	13	2	<1	2

Table 12: This table presents the carbon emissions (in g) of the inference on the test set of each model over each dataset.