PharmaER.IT: an Italian Dataset for Entity Recognition in the Pharmaceutical Domain

Andrea Zugarini[1,†], Leonardo Rigutini[1,*,†]

[1] expert.ai, Siena (Italy)

Abstract

Despite significant advances in Natural Language Processing, applying state-of-the-art models to real-world business remains challenging. A key obstacle is the mismatch between widely used academic benchmarks and the noisy, imbalanced data often encountered in domains such as finance, law, and medicine, especially in non-English languages, where resources are typically scarce. To address this gap, we introduce *PharmaER.IT*, a new dataset for entity recognition in the pharmaceutical and medical domain for the Italian language. *PharmaER.IT* is constructed from drug information leaflets obtained from the Agenzia Italiana del Farmaco, and annotated using either semi-automatic or fully automatic methods. The dataset comprises two complementary corpora: (1) the GOLD corpus, consisting of 57 leaflets annotated via a committee-based algorithm followed by expert manual validation, yielding 16833 high-quality entity mentions; and (2) the SILVER corpus, containing 2138 leaflets annotated solely through the automatic pipeline, without any human curation. We establish reference performance evaluating a range of token classification models and several LLMs under zero-shot conditions.

Keywords

NER, Pharmaceutical NER, Dataset, LLM

1. Introduction

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP), and arguably among the most demanded in industrial applications. While recent advances in transformer-based models [1] and Large Language Models (LLMs) [2, 3, 4, 5] have significantly improved entity extraction performance on standard benchmarks, their application to real-world business and professional contexts remains difficult. A primary challenge is the discrepancy between academic datasets and the often noisy, domain-specific texts encountered in real-world practice. This challenge is further exacerbated in specialized domains such as finance, law, and medicine. When dealing with languages other than English, annotated resources are even scarcer or non-existent.

In the medical and pharmaceutical domain, accurate entity recognition is critical for applications ranging from drug safety monitoring to automated clinical documentation. However, the Italian language remains underrepresented in the landscape of medical NER resources, limiting the development and evaluation of robust systems for local healthcare and regulatory contexts. Existing datasets are either too small, lack sufficient domain specificity, or are unavailable for public use due to privacy or licensing restrictions.

In this article we present *PharmaER.IT*, a novel dataset for NER in the pharmaceutical field for the Italian language. The dataset is derived from Riassunti delle Caratteristiche del Prodotto (RCPs), the official drug information leaflets, made publicly available by the Agenzia Italiana del Farmaco (AIFA).

PharmaER.IT is composed of two complementary corpora: a curated GOLD corpus, consisting of 57 RCPs annotated using a committee-based approach and refined through expert manual validation, and a SILVER corpus, comprising 2138 RCPs automatically annotated without human intervention.

This dual-corpus structure enables both evaluation and large-scale experimentation, facilitating the development of high-quality models as well as scalable weakly supervised approaches. To establish baseline performance, we evaluate a range of token classification models and several LLMs under diverse zero-shot settings.

2. Related work

CoNLL-2003 [6], was one of the first NER datasets and it is still a reference corpus for NER. It was constituted of news articles annotated with four entity types: person (PER), organization (ORG), location (LOC) and miscellaneous (MISC). Recently, numerous NER datasets have been released, many of which have been constructed using semi-automatic or fully automated annotation methods [7, 8, 2, 9], significantly expanding the entity tag-set. For instance, In Pile-NER [2], annotations were distilled from ChatGPT, resulting in about 45 thousands examples in English and more than 13 thousands distinct entity

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

[†] These authors contributed equally.

azugarini@expert.ai (A. Zugarini); lrigutini@expert.ai (L. Rigutini)

0000-0003-0344-1656 (A. Zugarini); 0000-0002-6309-2542 (L. Rigutini)

URL 1
The URL scheme of the RCPs on the AIFA website.

```
https://api.aifa.gov.it/aifa-bdf-eif-be/1.0.0/organizzazione/[sis]/farmaci/[aic]/stampati?ts=RCP
```

types. Even so, these resources do not target documents from vertical domains, such as finance or health. Other works proposed domain-specific NER corpora, such as in the financial [10, 11] and healthcare [12, 13] domains. However, these datasets are in English and mainly consist of well-curated, isolated sentences. In contrast, our work focuses on technical descriptions of pharmaceutical drugs.

Italian NER Datasets. The availability of NER data sets in Italian is extremely limited, particularly outside the traditional general-purpose domains and entity labels set [14]. Indeed, most NER datasets focus on news and social media contents [15, 16, 17, 18, 19].

Recently, it was introduced Multinerd [20], a multilingual dataset, covering Italian and a set of 15 distinct entity types. Among them, *Disease* and *Biological Entity* classes were included. However, examples originated from Wikipedia and Wikinews sentences, which are typically educational and encyclopedic. In *PharmaER.IT* instead, we collected drug leaflets, which present an highly technical and specialized lexicon. As an alternative strategy, [21] proposed to translate existing healthcare English datasets for NER, in Italian. Nonetheless, automatic translation may introduce errors or segmentation issues, especially on such a vertical domain.

3. Data Collection

In order to create a highly pharmaceutical-oriented dataset in Italian, we collected documents from the AIFA website.

Target documents

The AIFA is the official government institution that regulates the distribution of drugs in Italy[1]. The agency maintains the list of drugs authorized for sale in Italy, the list of pharmaceutical companies producing them and all the documentation made available by the manufacturer for each drug, including the drug leaflet. The leaflet is the short information document that accompanies the drug in the package and is divided into two types:

- Foglietto Illustrativo (FI) - the Package Leaflet. This is a document aimed at patients, with a simplified structure and language.
- Riassunto delle Caratteristiche del Prodotto (RCP). RCPs are documents for healthcare professionals, with a slightly more complex structure and more technical language and content. RCPs are approved documents, part of the marketing authorization for a drug, and intended primarily for healthcare professionals, adopting a medical-scientific terminology. In particular, RCPs contain detailed information on how to use the medicine, for example: therapeutic indications (what the medicine treats), dosage and method of administration, contraindications, special warnings, mechanism of action, side effects.

Given the strong technical content, we chose to use RCPs to build our data set. We choose to use only RCPs and ignore the FI since (1) the contents of the FIs are a subset of the contents of the second, and (2) the RCPs contain technical information relating to pharmacological properties and therapeutic indications that provide information of diagnostic-prescriptive value.

Data download

For each drug authorized for sale in Italy, AIFA assigns the unique AIC (Authorization for Placing on the Market) identification code[2] consisting of 9 digits in which the the 3 most significant on the left identify the type of packaging (capsules or syrup, mg, etc.), while the remaining 6 on the right (eventually padded with zeros) uniquely identify the drug (it is also referred as AIC6). Similarly, also for the companies producing the authorized drugs, AIFA assigns an unique three-digit code called SIS.

The open-data section of the AIFA website[3] contains several databases, including the list of drugs approved and distributed in Italy. This list can be downloaded as a csv file[4] and itemizes the AIC codes of the authorized class A and class H drugs, together with a series of auxiliary information. In addition, the AIFA website also provides an API endpoint to download all available documentation for authorized drugs. The API allows to download a drug RCP by specifying the SIS and the AIC6 codes and the type of document required (RCP) in the request URL, according to the scheme reported in URL 1.

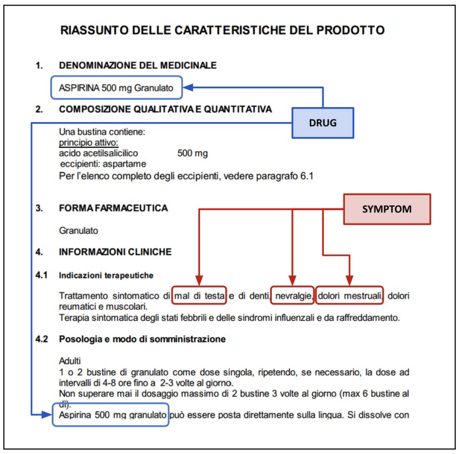
[1] https://www.aifa.gov.it/

[2] In collaboration with the European Medicines Agency (EMA), if the drug is intended for multiple European countries

[3] https://www.aifa.gov.it/open-data

[4] https://www.aifa.gov.it/web/guest/liste-dei-farmaci

Using the drug list and the download API, we collected 8634 RCP files in PDF format relating to class A and class H drugs, which were then converted to raw text files using a PDF-to-text conversion tool. Figure 1 shows an excerpt from the first page of a downloaded RCP.



RIASSUNTO DELLE CARATTERISTICHE DEL PRODOTTO

1. DENOMINAZIONE DEL MEDICINALE

ASPIRINA 500 mg Granulato

2. COMPOSIZIONE QUALITATIVA E QUANTITATIVA

Una bustina contiene:
principio attivo:
acido acetilsalicilico 500 mg
eccipienti: aspartame
Per l'elenco completo degli eccipienti, vedere paragrafo 6.1

3. FORMA FARMACEUTICA

Granulato

4. INFORMAZIONI CLINICHE

4.1 Indicazioni terapeutiche

Trattamento sintomatico di mal di testa e di denti, nevralgie, dolori mestruali, dolori reumatici e muscolari.
Terapia sintomatica degli stati febbrili e delle sindromi influenzali e da raffreddamento.

4.2 Posologia e modo di somministrazione

Adulti
1 o 2 bustine di granulato come dose singola, ripetendo, se necessario, la dose ad intervalli di 4-8 ore fino a 2-3 volte al giorno.
Non superare mai il dosaggio massimo di 2 bustine 3 volte al giorno (max 6 bustine al dì).
Aspirina 500 mg granulato può essere posta direttamente sulla lingua. Si dissolve con

Figure 1: Example of Italian RCP downloaded from AIFA with some entities highlighted (part of the first page).

4. Data annotation

For data labeling, we followed a semi-automatic procedure which included a human in the loop. Specifically, we first exploited a "committee" approach based on the use of two different automatic annotation models. Secondly, the annotated documents were reviewed by humans, with particular attention to the cases of discordant annotations returned by the two automatic annotators.

Tag-set

In designing the tag-set, we identified three families of data points: *Chemicals* , *Condition* and *Organism*. In this way, for each family it is possible to define a subset of related entity types. In the first version of the data, only *Condition* has more than an entity type, but we intend to extend the groups in the future. The resulting tag-set is reported in Table 1.

Automatic Pre-annotation

In the first step of annotation, documents are automatically labeled using a committee approach. The idea is to employ a limited group of automatic annotators, usually consisting of different algorithms and models, to generate multiple annotations for a single file. The acceptability of each annotation is subsequently assessed by examining the levels of concordance and discordance among these automatic annotators. We selected two approaches that were considered very different so that the concordance cases would provide a higher degree of reliability: (a) a neural annotator based on the use of a generative LLM and (b) a symbolic pre-annotator based on the use of an NLP Platform.

LLM-based pre-annotator. This automatic annotator was based on the use of a generative LLM. In particular, using a prompt specifically designed and developed for this task (Prompt 1), for each data-point type, this annotator model asks to the LLM to identify all the entities present within the text (provided as input) belonging to the target data-point. When the length of the content exceeds the input size of the LLM, the content of the drug information leaflet is divided into smaller chunks respecting the sentence grammar. We chose to use LLama 3.1 70B[5], a state-of-the-art, open source generative language model released by Meta that has reported excellent performance on several NLP tasks.
Given the generative nature of the LLM (as opposed to the word-classifying nature of the task), the result consists of a list of the identified entities without the position in which they were found (the start-end pairs). To obtain the final set of occurrences, a post-processing procedure performs the placement of the entities in the text using a string-matching search approach.

Rule-based pre-annotator. This annotator was based on the use of deep linguistic analysis. In particular, we used a NLP platform that, thanks to integrated linguistic resources (knowledge graph, semantic disambiguator and linguistic rules), allowed us to identify the occurrences of entities within the RCPs. For this task we used the proprietary NLP Platform of expert.ai[6] which consists in an integrated environment for deep language understanding and provides a complete natural language workflow with end-to-end support for annotation, labeling, model training, testing and workflow orchestration. To increase the recognition performances, the selected NLP Platform has been specialized by integrating knowledge and linguistic rules for the medical and pharmaceutical domains.
Given the lower generalization capacity typical of expert system approaches, for the entities identified by the rule-based annotator but missed by the LLM-based annotator, an additional verification step was included. In particular, for such cases of discordance, a further query was performed to an LLM in which confirmation of the extraction performed by the linguistic annotator is requested. The used prompt is reported in Prompt 2. For this additional step, we used the OpenAI GPT APIs[7], and

[5] https://ai.meta.com/blog/meta-llama-3-1/?utm_source=chatgpt.com
[6] https://www.expert.ai/products/expert-ai-platform/
[7] https://openai.com/api/

Table 1
The tag-set defined in *PharmaER.IT*.

Tag Family	Tag Name	Description
Chemicals	*DRUGS*	A synthetic or natural substance with beneficial effects in treating diseases, including product names and drug types.
Condition	*DISEASES*	Any pathological state or alteration of the body or one of its organs, including malformations, mental disorders, and injuries.
	SYMPTOMS	A morbid event indicating the presence of a disease.
Organism	*ANATOMICAL_PARTS*	All human body parts anatomically described, including organs, limbs, and anatomical structures.

Prompt 1
The prompt used for the LLM-based pre-annotator.

```
You are an expert in the field of pharmacology.

### INSTRUCTIONS

You are provided with a portion of a drug leaflet in Italian.
Your task is to identify and extract relevant entities regarding:

{TAG NAME with description}

Report the terms and expressions in singular form.
Branches of medicine such as 'pulmonology', 'traumatology', etc. are NOT diseases.
Return the result in a structured JSON.
Be sure to include only the terms 'explicitly' mentioned in the leaflet.
Exclude any interpretation or inference outside the text provided.
Report the terms as they are mentioned in the text, DO NOT make any changes or normalizations.

### TEXT

{Drug leaflet content}
```

in particular the "gpt-4o-mini" model.

Human Review

The output of the automatic pre-annotation phase consists of duplicate versions of the same RCP, each with labels inserted by the two different automatic annotators. To produce the single and final annotated version, a subsequent review phase was necessary in which, for each document, the outputs of the two models were analyzed in order to be accepted or rejected, and in which any tags missed during the pre-annotation phase could also be inserted. In particular, a merged version of each RCP was then created, reporting the outputs of the two annotators, also highlighting the cases of agreement (both models had identified the occurrence of an entity) and disagreement (only one of the two models had hypothesized the occurrence of an entity). These "merged" documents were then distributed to human experts to examine the annotations inserted by the pre-annotation phase (accepting or rejecting them) and with the possibility of adding new ones.

For this human validation phase, we employed a panel of five human reviewers and assigned each of them a set of RCPs randomly drawn from the total. To subsequently measure the degree of consistency of the final annotation outputs, we designed the assignment so that part of them were blindly shared between two reviewers. In this way, we obtained a total of 57 RCPs selected for human review, 6 of which were randomly and hiddenly assigned to two reviewers. This step has been performed using the annotating support of the expert.ai natural language platform[6].

Review guidelines. To improve the consistency of the final annotations, a document containing guidelines was drafted and provided to the reviewers. In this document, a set of indications on how to consider ambiguous cases were specified, mainly based on the context in which they appear.

Annotation quality assessment

To estimate the quality of the final annotation outputs, we exploited the set of 6 RCPs reviewed by a pair of human experts. In particular, by indicating with $L_1(RCP_i)$ and $L_2(RCP_i)$ the two sets of annotations resulting from the review phase of reviewer rev_1 and rev_2 respectively

Prompt 2
The prompt used for the LLM-based validation step for the linguistic-based pre-annotator.

```
You are an expert in the field of pharmacology.

Check if the word {WORD} provided in the following text could be a {TAG NAME}

Text: {TEXT}

Answer only with YES or NO.
No preamble or anything else.
```

on the same RCP_i document, we calculated several standard indices[8]:

(a) Joint Probability of Agreement, which measures the chance of having a match between the annotations resulting from the two reviewers: $\mathit{JPA} = \frac{\#(L_1 \cap L_2)}{\#(L_1 \cup L_2)}$.

(b) Conditional Probability of Agreement of rev_k, which measures the naive probability that annotations resulting from the reviewer k have a match with the annotations resulting from the other reviewer: $\mathit{CPA} = \frac{\#(L_1 \cap L_2)}{\#(L_k)}$, $k \in \{1, 2\}$.

(c) Coverage of rev_k, which measures the probability that a randomly selected annotation in RCP_i comes from the reviewer k: $\mathit{Cov} = \frac{\#(L_k)}{\#(L_1 \cup L_2)}$, $k \in \{1, 2\}$.

(d) Cohen's kappa (κ), which extends the Joint Probability of Agreement taking into account that agreement may occur by chance [22]: $\kappa = \frac{p_o - p_e}{1 - p_e}$ where $p_o = \mathit{JPA}$ is the observed agreement, $p_e = \frac{\#(L_1) \times \#(L_2)}{N^2}$ estimates the probability of a random agreement and $N = \#(L_1 \cup L_2)$ is the total number of annotations.

The values were evaluated for each RCP_i, and then averaged over all RCPs (micro-average), separately for each data point.

Table 2
The quality assessment results of the output of the annotation and validation process.

Data point	***JPA***	***CPA***	***Cov***	κ
DRUGS	0.85	0.91	0.91	0.90
DISEASES	0.98	0.84	0.86	0.83
SYMPTOMS	0.74	0.86	0.87	0.84
ANATOMICAL_PARTS	0.68	0.84	0.84	0.76
Average	**0.81**	**0.86**	**0.87**	**0.83**

The results are reported in the Table 2 and the values of Cohen's kappa (κ) show a substantial agreement in the resulting annotated data [23].

5. The *PharmaER.IT* dataset

The resulting *PharmaER.IT* dataset consists of the two corpora: the GOLD corpus and the SILVER corpus. We made it publicly available and free-to-download from HuggingFace[9].

The GOLD corpus

The GOLD corpus consists of the 57 labeled RCPs which were annotated by the semi-automatic procedure described in Section 4. It is composed by a total of 16833 occurrences of entities identified by automatic pre-annotators and then reviewed by humans. We then established a predefined split for training, validation and testing, randomly selecting 37, 10 and 10 RCPs respectively. The resulting distribution of occurrences of the data points is reported in Table 3.

Table 3
Distribution of annotated entities in the GOLD corpus.

Data point	**Train**	**Validation**	**Test**	**Total**
DRUGS	5911	716	1222	7849
DISEASES	3344	477	614	4435
SYMPTOMS	2582	363	480	3425
ANATOMICAL_PARTS	817	121	186	1124
Total	12654	1677	2502	16833

The SILVER corpus

To build the SILVER corpus we sampled 2138 leaflets from the remaining 8567 documents. These RCPs were pre-annotated with algorithm in Section 4, without any revision from human annotators. The resulting documents were added to the *PharmaER.IT* dataset as a SILVER corpus. Table 4 shows the distribution of

[8] https://en.wikipedia.org/wiki/Inter-rater_reliability

[9] https://huggingface.co/datasets/expertai/PharmaER.IT

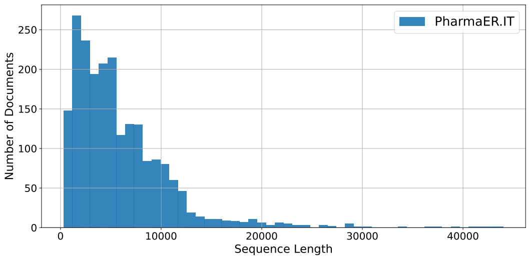


Figure 2: Sequence length distribution (in words) of *PharmaER.IT* documents in SILVER partition.

annotations within SILVER.

Table 4
Distribution of annotated entities in the SILVER corpus.

Data point	Total
DRUGS	385210
DISEASES	245240
SYMPTOMS	80763
ANATOMICAL_PARTS	70587
Total	781800

6. Experiments

NER has been traditionally tackled as a token classification problem, with models fine-tuned on the downstream task. With the emergence of LLMs, alternative approaches to NER based on prompting in zero-shot or few-shot settings have gained popularity. Therefore, we established a set of baselines on *PharmaER.IT* using both strategies and a wide range of models.

Experimental setup

Models. We evaluated several state-of-the-art transformer-based architectures for token classification that are widely adopted: bert [24], roberta [25] and xlm-roberta [26]. We studied them on different sizes and pre-trained versions specialized for Italian[10] or multilingual.

Concerning LLMs, we considered several backbones ranging from 7B to 24B parameters, i.e. in the small-medium size tier. We paid particular attention to models that were either pre-trained or further adapted for the Italian language, or that explicitly included Italian in their pre-training corpora. In particular, we assessed on *PharmaER.IT* Llama-3.1-8B, LLaMAntino-3-8B [27], Minerva-7B [28], Velvet-14B[11], Salamandra-7B [29], EuroLLM-9B [30] and Mistral-Small-3.1-24B[12]. We tested two different prompts. A simple one where the LLMs is asked to generate a structured JSON with the entity types as keys and the entities extracted as values. In the second prompt, a definition and some annotation guidelines are specified for each class. In this second evaluation, we also considered SLIMER-IT-8B [5], a fine-tuned version of LLaMAntino for zero-shot NER that follows the approach of [31]. Differently from the rest, SLIMER-IT-8B extracts one entity type at a time, thus each context is repeated 4 times.

Document Chunking. *PharmaER.IT* documents are characterized by their considerable length and dense presence of annotated entities, which poses specific challenges for NER models based on transformer architectures with fixed-length input windows. As shown in Figure 2, many documents exceed the standard maximum token limit (e.g., 512 tokens). Documents' size is also problematic for LLMs, which – despite supporting longer contexts – still face practical limits, especially when there are hundreds of entities per document. Therefore, documents are split in chunks. For encoders, we tokenize documents with their respective tokenizer. We set a maximum length of 512 and a window stride of 64. Conversely, for LLMs text is split in passages of sentences having at most 768 characters.

Training. Encoder models were fine-tuned on the train/validation/test split reported in Table 3. To augment the training data, we also added the Silver corpus in the train set, and we evaluated its impact on the performance. We kept in all the experiments the learning rate fixed to $5 \cdot 10^{-5}$, 8 epochs and early stopping with patience 3. Batch size was set to 16 in all the experiments without silver data, and 128 otherwise. Unlike encoder-based models, LLMs were used without fine-tuning, relying solely on zero-shot prompts for entity extraction.

Metrics. All the models were evaluated on the test set measuring the F1 score. However, due to the different chunking and the non-positional nature of generative models, LLMs and token classifiers were evaluated independently. We adopted the standard micro-F1 score (simply denoted as F1) for token classification models on their positional predictions. In LLMs instead, evaluation occurs at document-level. First, we collect in each passage all the unique text spans extracted per-class by the LLM, then we measure the F1 score against all the

[10] https://huggingface.co/dbmdz/bert-base-italian-cased

[11] https://huggingface.co/Almawave/Velvet-14B

[12] mistralai/Mistral-Small-3.1-24B-Instruct-2503

Table 5
Results of state-of-the-art encoders fine-tuned for token classification on *PharmaER.IT*.

Model	Use Silver	Precision	Recall	F1	Δ (F1)
roberta	True	77.32	**75.96**	76.64	**+7.44**
	False	66.86	71.71	69.20	
roberta-large	True	**79.10**	75.45	**77.23**	**+5.57**
	False	71.42	71.91	71.66	
xlm-roberta	True	78.01	**76.12**	77.05	**+8.33**
	False	66.16	71.49	68.72	
xlm-roberta-large	True	78.06	74.92	76.46	**+4.25**
	False	70.25	74.28	72.21	
bert-multilingual-cased	True	77.07	74.85	75.94	**+9.64**
	False	64.60	68.10	66.30	
bert-italian-cased	True	76.80	75.05	75.91	**+7.12**
	False	65.37	72.57	68.79	

unique target entities of the document, following the UniNER [2] implementation[13]. Please note that these two F1 scores are computed on fundamentally different values, and therefore they are not comparable.

Results

Token Classification. From Table 5, we can observe that F1 score varies from about 66 to 72 across all models when fine-tuned on the training set without silver corpus. Roberta architectures yield the best scores, in particular xlm-roberta-large that achieves the best result (in the no silver setting).

Impact of Silver Partition. The results, shown in Table 5, clearly demonstrate that augmenting the training set with pre-annotated (silver) documents significantly enhances model performance. All evaluated models benefit from this data augmentation, with improvements reaching up to 9.64 F1 points. Notably, smaller models gain the most from the additional data, effectively narrowing the performance gap between base and large architectures. As a result, the base versions of RoBERTa and XLM-RoBERTa emerge as the best and second-best performing models, respectively.

Off-the-shelf LLMs. Zero-shot entity extraction of pharmaceutical entities is a challenging task in such an unfamiliar domain. Albeit Mistral-small and LLama-based models achieve relevant scores, other LLMs like, Salamandra, Minerva and Velvet-14B, were not able to follow the provided instructions. Therefore, we reported in Table 6 the F1 scores of only the models that were able to extract entities.

Table 6
Zero-shot extraction with simple prompt.

Model	Precision	Recall	F1
Llama-3.1-8B	38.90	38.47	38.69
LLaMAntino-3-8B	40.20	55.36	46.58
EuroLLM-9B	43.13	16.65	24.02
Mistral-Small-3.1-24B	**43.61**	**61.90**	**51.17**

LLMs with Definition and Guidelines. When the prompt is enriched with entity type definition and annotation guidelines, all the LLMs generally improve their scores, with the exception of LlaMAntino, which registers a small flexion. In particular, all the models extracted some entities. This suggests that with appropriate prompt design there is room for improving these baselines. Results are presented in Table 7.

Table 7
LLMs with definition and guidelines.

Model	Precision	Recall	F1
Minerva-7B	9.69	3.08	4.68
salamandra-7b	13.84	11.59	12.62
Llama-3.1-8B	35.24	**57.09**	43.58
LLaMAntino-3-8B	34.39	54.99	42.31
SLIMER-IT-8B	54.60	32.18	40.50
EuroLLM-9B	25.20	38.72	30.53
Velvet-14B	**58.33**	4.32	8.04
Mistral-Small-3.1-24B	53.63	56.47	**55.02**

7. Conclusions and future works

In this work, we presented *PharmaER.IT*, an Entity Recognition dataset for the pharmaceutical domain in Italian language. *PharmaER.IT* was created from AIFA drug information leaflets. It includes two corpora: a curated

[13] https://github.com/universal-ner

GOLD corpus 57 of created semi-automatically, and the SILVER corpus, consisting of 2138 annotated RCPs without human intervention.

To establish comprehensive baselines, we assessed a selection of different NER models, both based on token-classification models and zero-shot extraction with LLMs. The resulting *PharmaER.IT* dataset has been released in HuggingFace[14].

In the future, we intend to extend *PharmaER.IT* in two directions. On one side, we plan to increase the amount of manually labeled data and to extend the labels set with more domain-specific tags. On the other hand, we aim to introduce relations between entities in order to extend the dataset to Relational Extraction.

Acknowledgements

We thank S. Ligabue, V. Masucci, M. Spagnolli and S. M. Marotta for their support in the data preparation and annotation process.

The work was partially funded by:

- "MAESTRO - Mitigare le Allucinazioni dei Large Language Models: ESTRazione di informazioni Ottimizzate" a project funded by Provincia Autonoma di Trento with the Lp 6/99 Art. 5:ricerca e sviluppo, PAT/RFS067-05/06/2024-0428372, CUP: C79J23001170001[15];
- "ReSpiRA - REplicabilità, SPIegabilità e Ragionamento", a project financed by FAIR, Affiliated to spoke no. 2, falling within the PNRR MUR programme, Mission 4, Component 2, Investment 1.3, D.D. No. 341 of 03/15/2022, Project PE0000013, CUP B43D22000900004[16];
- Villanova, a project financed by IPICEI-CIS, Prog. n. SA. 102519 - CUP B29J24000850005[17].

References

[1] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023. arXiv:2311.08526.

[2] W. Zhou, S. Zhang, Y. Gu, M. Chen, H. Poon, Universalner: Targeted distillation from large language models for open named entity recognition, arXiv preprint arXiv:2308.03279 (2023).

[3] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[4] O. Sainz, et al., Gollie: Annotation guidelines improve zero-shot information-extraction, 2024. arXiv:2310.03668.

[5] A. Zamai, L. Rigutini, M. Maggini, A. Zugarini, Slimer-it: Zero-shot ner on italian language, arXiv preprint arXiv:2409.15933 (2024).

[6] E. F. T. K. Sang, F. D. Meulder, 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, 2003.

[7] D. S. Menezes, P. Savarese, R. L. Milidiú, Building a massive corpus for named entity recognition using free open data sources, arxiv (2019). URL: https://arxiv.org/abs/1908.05758. arXiv:1908.05758.

[8] D. Alves, G. Thakkar, M. Tadić, Building and evaluating universal named-entity recognition english corpus, arxiv (2022). URL: https://arxiv.org/abs/2212.07162. arXiv:2212.07162.

[9] N. Ringland, X. Dai, B. Hachey, S. Karimi, C. Paris, J. R. Curran, Nne: A dataset for nested named entity recognition in english newswire, in: 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019. URL: https://aclanthology.org/P19-1510/.

[10] L. Loukas, M. Fergadiotis, I. Chalkidis, E. Spyropoulou, P. Malakasiotis, I. Androutsopoulos, G. Paliouras, Finer: Financial numeric entity recognition for xbrl tagging, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 4419–4431.

[11] A. Zugarini, A. Zamai, M. Ernandes, L. Rigutini, Buster: a 'business transaction entity recognition' dataset, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2023. URL: https://doi.org/10.18653/v1/2023.emnlp-industry.57. doi:10.18653/v1/2023.emnlp-industry.57.

[12] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, Biocreative v cdr task corpus: a resource for chemical disease relation extraction, Database 2016 (2016).

[13] C. Quirk, H. Poon, Distant supervision for relation extraction beyond the sentence boundary, arXiv preprint arXiv:1609.04873 (2016).

[14] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J. M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, Computer Standards & Interfaces 35 (2013) 482–489. URL: https://www.sciencedirect.com/science/

[14] https://huggingface.co/datasets/expertai/PharmaER.IT

[15] MAESTRO: https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/C79J23001170001

[16] RESPIRA: https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/B43D22000900004

[17] Villanova: https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/B29J24000850005

article/pii/S0920548912001080. doi:https://doi.org/10.1016/j.csi.2012.09.004.

[15] C. Bosco, V. Lombardo, L. Vassallo, A. Lesmo, Building a treebank for italian: a data-driven annotation schema, in: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), 2000.

[16] B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, R. Sprugnoli, I-CAB: the Italian content annotation bank, in: N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, D. Tapias (Eds.), Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/518_pdf.pdf.

[17] V. Bartalesi Lenzi, M. Speranza, R. Sprugnoli, Named entity recognition on transcribed broadcast news at evalita 2011, in: B. Magnini, F. Cutugno, M. Falcone, E. Pianta (Eds.), Evaluation of Natural Language and Speech Tools for Italian, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 86–97.

[18] P. Basile, A. Caputo, A. Gentile, G. Rizzo, Overview of the evalita 2016 named entity recognition and linking in italian tweets (neel-it) task, 2016.

[19] T. Paccosi, A. Palmero Aprosio, KIND: an Italian multi-domain dataset for named entity recognition, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 501–507. URL: https://aclanthology.org/2022.lrec-1.52.

[20] S. Tedeschi, R. Navigli, MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation), in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 801–812. URL: https://aclanthology.org/2022.findings-naacl.60. doi:10.18653/v1/2022.findings-naacl.60.

[21] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localizing in-domain adaptation of transformer-based biomedical language models, Journal of Biomedical Informatics (2023). URL: https://doi.org/10.1016/j.jbi.2023.104431. doi:10.1016/j.jbi.2023.104431.

[22] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37 – 46.

[23] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[26] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[27] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, arXiv preprint arXiv:2312.09993 (2023).

[28] R. Orlando, L. Moroni, P.-L. H. Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva llms: The first family of large language models trained from scratch on italian data, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 707–719.

[29] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. Da Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, et al., Salamandra technical report, arXiv preprint arXiv:2502.08489 (2025).

[30] P. H. Martins, J. Alves, P. Fernandes, N. M. Guerreiro, R. Rei, A. Farajian, M. Klimaszewski, D. M. Alves, J. Pombal, M. Faysse, et al., Eurollm-9b: Technical report, arXiv preprint arXiv:2506.04079 (2025).

[31] A. Zamai, A. Zugarini, L. Rigutini, M. Ernandes, M. Maggini, Show less, instruct more: Enriching prompts with definitions and guidelines for zero-shot ner, arXiv preprint arXiv:2407.01272 (2024).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) in order to: Paraphrase and reword, Improve writing style, and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.