Show Less, Instruct More: Enriching Prompts with Definitions and Guidelines for Zero-Shot NER

Anonymous ACL submission

Abstract

Recently, several specialized instruction-tuned Large Language Models (LLMs) for Named Entity Recognition (NER) have emerged. Compared to traditional NER approaches, these models have strong generalization capabilities. Existing LLMs mainly focus on zero-shot NER in out-of-domain distributions, being fine-tuned on an extensive number of entity classes that often highly or completely overlap with test sets. In this work instead, we propose SLIMER, an approach designed to tackle never-seen-before named entity tags by instructing the model on fewer examples, and by leveraging a prompt enriched with *definition* and *guidelines*. Experiments demonstrate that definition and guidelines yield better performance, faster and more robust learning, particularly when labelling unseen Named Entities. Furthermore, SLIMER performs comparably to state-of-the-art approaches in out-of-domain zero-shot NER, while being trained on a reduced tag set.

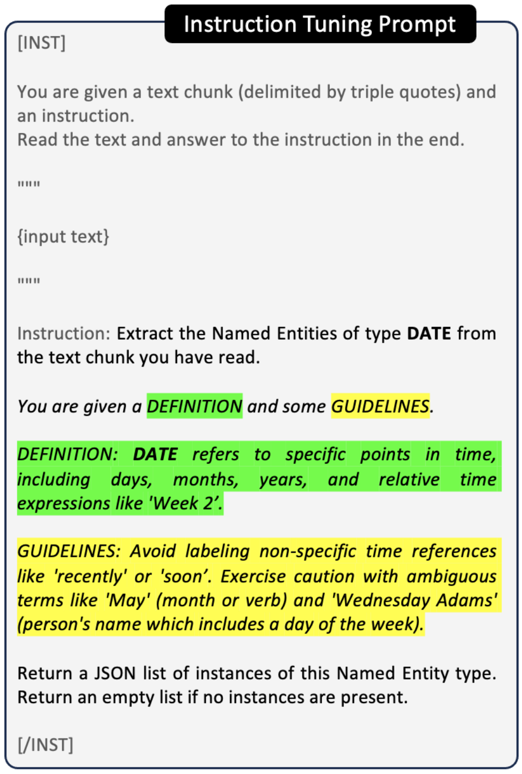

Figure 1: SLIMER's instruction-tuning prompt. Dedicated entity definition and guidelines steer the model generation.

1 Introduction

Named Entity Recognition (NER) is a crucial problem in Natural Language Processing (NLP), usually being a key component in Information Extraction pipelines. Traditional methods frame NER into a sequence labeling task (Li et al., 2020), where models are specialized on a narrow domain and a pre-defined label set, thus lacking generalization capabilities outside the downstream task at hand.

On the contrary, Large Language Models (LLMs) have demonstrated strong zero-shot capabilities. Through the use of cleverly designed prompts, models like GPT-3 can tackle NER via In-Context Learning (Radford et al., 2019; Brown et al., 2020). However, smaller (encoder-only) LMs trained on specific NER tasks may still outperform LLMs (Wang et al., 2023a; Ye et al., 2023; Zhou et al., 2023). To this end, Instruction-Tuning of LLMs has emerged as an effective method to improve their performance (Wei et al., 2022; Chung et al., 2022; Wang et al., 2022b). There exist several works in the literature based on instruction-tuning for NER, such as InstructUIE (Wang et al., 2023b), UniNER (Zhou et al., 2023), GoLLIE (Sainz et al., 2024) and GNER (Ding et al., 2024).

Such dedicated LLMs have the ability to perform zero-shot NER on heterogeneous input domains and a multitude of possibly never-seen-before NEs. Existing works mainly focus on zero-shot NER in Out-Of-Distribution (OOD) input domains, while fine-tuning on an extensive number of entity classes that often highly or completely overlap between the training and test sets. Consequently, the problem

	InstructUIE	UniNER	GoLLIE	GLiNER	GNER	SLIMER (our)
Architecture	enc-dec	decoder	decoder	encoder	(enc-)dec	decoder
NER-only	×	✓	×	✓	✓	✓
nested-NER	✓	✓	✓	×	×	✓
Instruction-tuned	✓	✓	✓	×	✓	✓
Instruction template	list tuples	conversation	Py-classes	×	gen-BIO	guided inst
NE Guidelines	×	×	✓	×	×	✓
Inference cost	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{X}\rvert \times \lvert\mathcal{Y}\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{X}\rvert \times \lvert\mathcal{Y}\rvert$
Works document level	✓	✓	✓	✓	×	✓
Trained on synthetic data	×	✓	×	✓	✓	✓
# distinct NEs	≤ 119	13020	≤ 40	13020	13020	391
Human effort for guidelines	×	×	high	×	×	gpt-prompt
Out-Of-Domain evaluation	✓	✓	✓	✓	✓	✓
Unseen NEs evaluation	×	×	✓	×	×	✓

Table 1: Overview of existing works in the literature, highlighting the differences on some identified key comparative features. $|\mathcal{X}|$ denotes the number of text inputs, $|\mathcal{Y}|$ the number of NEs. The symbol $\leq$ is used to indicate the upper bound on the number of distinct NEs in training when there is no overlap of label sets between the merged datasets.

of tagging unseen named entities has been little investigated, with GoLLIE as the only exception.

In this work instead, we tackle both scenarios by carefully selecting a training set with limited degree of overlap with test data. To facilitate effective zero-shot capabilities on such novel NEs, similar to GoLLIE, we steer the model with annotation guidelines. Unlike GoLLIE, we drop their code-based representation in favour of a more natural instruction that includes a *definition* and some *guidelines* for the category being extracted. This simplifies the prompt design and it allows us to automatically generate synthetic definition and guidelines by means of another LLM.

We study how increasing both the number of training samples and the number of unique Named Entity (NE) categories affect the generalization capabilities of LLMs, with or without the support of definitions and guidelines. By using fewer training samples from a reduced number of distinct named entity tags, combined with prompts enriched of *definition* and *guidelines*, we name our approach **SLIMER**: **S**how **L**ess, **I**nstruct **M**ore - **E**ntity **R**ecognition[1].

Experiments were conducted on two standard NER benchmarks for OOD, MIT (Liu et al., 2013) and CrossNER (Liu et al., 2021). Additionally, we assessed performance on never-seen-before NEs on BUSTER (Zugarini et al., 2023), a document-level NER dataset with entity tags novel to all the evaluated models.

Comparison of SLIMER with its baseline, i.e. the model devoid of definition and guidelines, reveals SLIMER's deeper understanding, faster and more stable learning, and better zero-shot performance. Despite being trained on a fraction of the data, with little overlap between train and test named entity tags, SLIMER performs comparably against state-of-the-art instruction-tuning approaches, revealing stronger generalization capabilities when dealing with unseen named entities.

2 Related Work

Commonly employed machine-learning solutions frame NER into a "sequence labeling task", where the goal is to assign a BIO label to each element in a given sequence (Li et al., 2020). Fine-tuning BERT-family models (Devlin et al., 2018) for NER is a well established approach. While these models excel in supervised contexts, they have the severe limitation of being constrained to a predefined set of labels and inputs from limited domains, making it difficult to generalize across different contexts and on unseen named entities.

2.1 In-Context Learning

Radford et al. (2019) were among the first to explore the ability of LLMs to perform "multi-task learning". In this setting, tasks are formulated as text-to-text problems and natural language instructions are prefixed in the input to prompt the model towards the task it has to solve. Building on this breakthrough, Brown et al. (2020) enhanced these zero-shot and few-shot multi-task capabilities, paving the way for what has been termed "In-Context Learning" or "Prompt Engineering".

While LLMs have demonstrated impressive capabilities in zero-shot settings on various challenging tasks, endeavours to utilize LLMs for Informa-

[1]SLIMER will be made publicly available.

MIT		CrossNER					BUSTER	TOT
Movie	Restaurant	AI	Literature	Music	Politics	Science		
$0/12$ 0%	$1/8$ 13%	$4/13$ 31%	$4/11$ 36%	$4/12$ 33%	$4/8$ 50%	$4/16$ 25%	$0/6$ 0%	$21/86$ 24%

Table 2: Overlap (%) between the NEs seen by SLIMER in training and those present in MIT and CrossNER benchmarks. For UniNER, GLINER and GNER models this overlap is 100%, i.e. there are no unseen NEs in test. The NEs in BUSTER are novel to all models.

tion Extraction have been less promising (Keraghel et al., 2024). Attempts to make use of LLMs (such as GPT) through clever prompt engineering have been conducted by Wang et al. (2023a) in their paper GPT-NER. In Ye et al. (2023), the authors compared several GPT models on various NLU tasks, including NER. The results highlighted a significant gap compared to supervised encoder-only approaches. Those results were further confirmed by Zhou et al. (2023) on other NER datasets.

2.2 Fine-Tuning for Zero-Shot NER

Parallel to prompt engineering LLMs, other approaches explored the way of instruction-tuning LLMs (Wei et al., 2022; Chung et al., 2022; Wang et al., 2022b), often involving models of smaller size and focusing the training on a single task, for example NER. InstructUIE (Wang et al., 2023b) is a encoder-decoder T5-11B model fine-tuned on supervised IE datasets (among which NER) phrased as text-to-text problems, where the instruction exhorts the model to return a list of tuples *(text span, entity type)*, choosing between a provided list of categories. UniNER (Zhou et al., 2023) consists in a decoder-only LLaMA model finetuned on a "conversation style template". In inference, the model is prompted with the question "What describes NE in the text?" and a list of text spans that belong to the requested NE category is returned.

Based on this instruction-tuning approach, several other works have emerged, each based on a different instruction-tuning template, with the aim of further improving the performance of LLMs in both supervised and zero-shot NER. Table 1 provides an overview of some of the most significant current approaches, highlighting the adopted backbone architecture, the designed instruction-tuning template, and other key comparative features.

GoLLIE (Sainz et al., 2024) is the first (and currently the only) that includes annotation guidelines in its prompt. In particular, the authors adopted a code-based representation encoding the NE labels as Python classes and providing the guidelines as doc-strings. So far, GoLLIE is the only approach that was also evaluated on unseen named entities tagging (see Table 1).

GNER (Ding et al., 2024) rethinks the importance of negative instances (i.e., "O" tags in BIO labeling) and replaces the established entity-centric schema with a BIO-like generation, replicating the same input text along with token-by-token BIO labels. Despite the limitations on the input length and output parsing difficulties, their approach displays stronger boundaries detection and reduced classification indecision compared to the other approaches.

Finally, GLiNER (Zaratiana et al., 2023) consists of a much smaller model based on an encoder-only and non-instruction-tuned architecture, which achieves a remarkable performance in both supervised and zero-shot NER.

3 SLIMER

This section presents our approach, SLIMER. First, we provide motivations for reducing the number of training samples. Then, we describe the instruction-tuning prompt. Finally, we discuss how to generate definitions and guidelines automatically by means of another LLM, such as ChatGPT.

3.1 Show Less

Existing models for zero-shot NER are trained on a large set of entity tags and examples. This training data can be generated synthetically (Zhou et al., 2023), by merging existing human-labelled datasets (Sainz et al., 2024), or even combing the two approaches (Zhou et al., 2023). While training on such extensive data certainly strengthens cross-domain zero-shot NER performance, it is unclear how it affects generalization capabilities on never-seen-before entity tags. Furthermore, as already observed in literature, instruction tuning helps aligning the model to the task and desired output format, but most of the gains in performance are achievable with little amounts of instructions (Wang et al., 2022a; Zhou et al., 2024; Zugarini et al., 2024). Motivated by this, we train SLIMER on a fraction of the training data that is typically used to instruction-tune zero-shot models for NER.

3.2 Instruct More

As we reduce the data, in contrast, we enrich the model prompt with a *definition* and some *guidelines* about the entity to tag. An example of instruction tuning prompt is illustrated in Figure 1.

Definition and Guidelines. The definition for a NE is designed to be a short sentence describing the tag at hand. This is followed by guidelines that provide annotation directives to align the model's labeling with the desired annotation scheme. Guidelines can be used to discourage the model from labelling particular edge cases or to provide examples of such NE. Thus, these components are intended to better instruct the model with specifics that well define what to extract and what not to extract. Moreover, such an information is crucial when dealing with unfamiliar entity tags, and it also allows to distinguish between polysemous categories. From now on, we refer to Definition and Guidelines as D&G.

Prompt structure. According to Zhou et al. (2023), we designed the prompt to extract the occurrences of one entity type per call. This has the drawback of requiring |NE| inference calls on each input text, but allows the model to better focus on a single NE type at the time. Moreover, compared to GoLLIE, where all tags's guidelines are prepended to the input, each individual instruction will be simpler and shorter.

Output. We ask the model to generate its output in a parsable JSON format consisting of a list of NE instances identified in the given input. It is worth noticing that, in the LLM fine-tuning the Next To Prediction-loss penalizes the order of the returned tokens. Hence, during training we sort the target entities by their order of appearance within the input text. Moreover, since it's redundant to return the same instance text multiple times, we reduce the list of target instances to a set of unique text instances.

3.3 Definition and Guidelines generation

To fully exploit the potential benefits of D&G, we must have high-quality information about an entity.

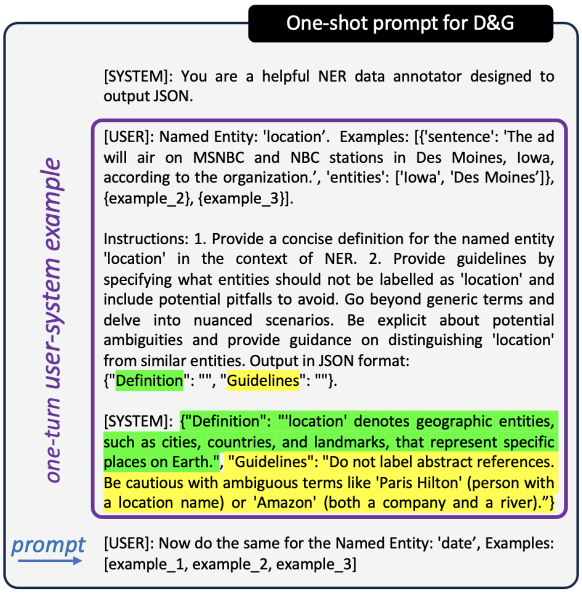

Figure 2: Prompt for generating the Definition and the Guidelines for a specific named entity.

When the number of entity types is small, their production can be tackled manually, but as the number grows, it may require excessive human effort, as also pointed out in GoLLIE (Sainz et al., 2024).

To overcome this limitation, we exploited the OpenAI's Chat-GPT APIs to automatically generate definition and guidelens. In particular, we designed the *one-shot prompt template* reported in Figure 2, which was used to query gpt-3.5-turbo-1106. An exemplary one-round user-system conversation is used to illustrate the desired output to the model. The three examples are randomly sampled for each NE from the dataset at hand. Thanks to such a prompt, all the generated definitions and guidelines exhibit a similar structure, with short and clear defining sentences, and with guidelines highlighting edge cases where to be cautious.

4 Experiments

In the experiments, we investigate the impact of definition and guidelines on zero-shot NER. We compare SLIMER against state-of-the-art models on out-of-distribution input domains and unseen entity types. Furthermore, we study how increasing the number of training samples affects performance.

4.1 Datasets

The LLMs have all been fine-tuned on a subset of PileNER-type (Zhou et al., 2023), and evaluated on three different benchmarks: MIT (Liu et al., 2013) and CrossNER (Liu et al., 2021) to assess out-of-distribution (OOD) performance, BUSTER (Zugarini et al., 2023) to measure performance on never-seen-before Named Entities.

PileNER-type. PileNER-type (Zhou et al., 2023) is a synthetic dataset comprising a large set of approximately 50 thousands examples, encompassing over 13 thousands different named entity types. We kept only a subset of them, considering only those NEs with at least 100 instances. From the remaining 455 named entity classes, we manually revised and merged together classes of identical types spelled differently, e.g. *organisation* and *organization*, and discarded some "catch-all" gpt-hallucinated labels (e.g. *unknown, other, miscellaneous, general, entity type*), thus further reducing to 423 different labels. Finally, to limit the overlap between training and test entity types, we excluded nearly all the categories present in the test datasets, with the exception of the standard NER tags: *person, location, organization, country*. Overall, we kept 391 distinct NEs. Percentages of overlap between train and test are reported in Table 2. The dataset composition, grouped by topic, is illustrated in Figure 3.

MIT and CrossNER. MIT (Liu et al., 2013) and CrossNER (Liu et al., 2021) datasets have become de-facto the standard benchmark for zero-shot NER. We use them to compare SLIMER against existing state-of-the-art models on out-of-distribution (OOD) domains.

BUSTER. We extend the evaluation beyond the MIT and CrossNER benchmarks by including BUSTER (Zugarini et al., 2023). The dataset is a document-level NER benchmark in financial domain. Both domain and named entity tags differ from what observed by all the models during instruction-tuning. Its significant differences from standard NER datasets make BUSTER a perfect benchmark for evaluating zero-shot performance on never-seen-before NEs.

4.2 Settings

Backbone and Training setup. SLIMER is based on LLaMA-2 7B chat (Touvron et al., 2023). Investigating how different families or model sizes affect results is outside the objectives of our work. In all the experiments the models were fine-tuned with LoRA (Hu et al., 2021) $r = 8$, $\alpha = 16$ for 10 epochs with early stopping, 32 batch size and

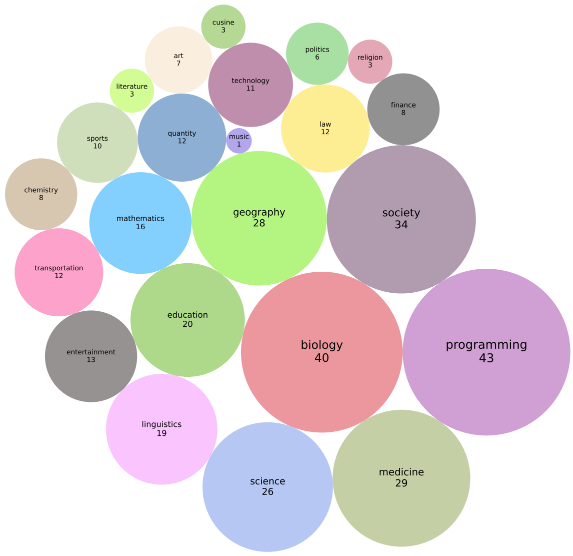

Figure 3: The 391 Named Entities in the PileNER-type subset, grouped by macro topics. "misc" (not shown) groups the 26 NEs that do not fit into the defined topics.

learning rate initialized to 3.0×10^{-4} with cosine scheduler and a warm-up of 60 steps. Context length was set to 768, longer inputs were chunked.

Metrics. We align with existing work in the literature by computing the F1 metric in a strict manner, i.e. for a given NE type, all the unique text spans within the text passage are required to be retrieved, each with all its associated tokens and no additional ones added. Any reported SLIMER result is the averaged value of three different runs of the same training-evaluation configuration. Hence, we also provide the standard deviation of our experiments.

4.3 Compared Models

We compare SLIMER to several state-of-the-art approaches for zero-shot NER:

ChatGPT (Radford et al., 2019), prompted with the same strategy in Ye et al. (2023). It constitutes a baseline not specifically instructed for NER.

InstructUIE (Wang et al., 2023b), based on the flan-t5-xxl encoder-decoder model with 11B parameters.

UniNER (Zhou et al., 2023), a family of LLMs all based on the LLaMA-1-7B architecture. We evaluate the three variants: *type*, *type+sup* and *def*. The first is trained on full PileNER-type, described in Subsection 4.1. Additionally, *type+sup.* was trained also on a collection of human-labeled NER datasets. UniNER-*def* instead, distinguishes from

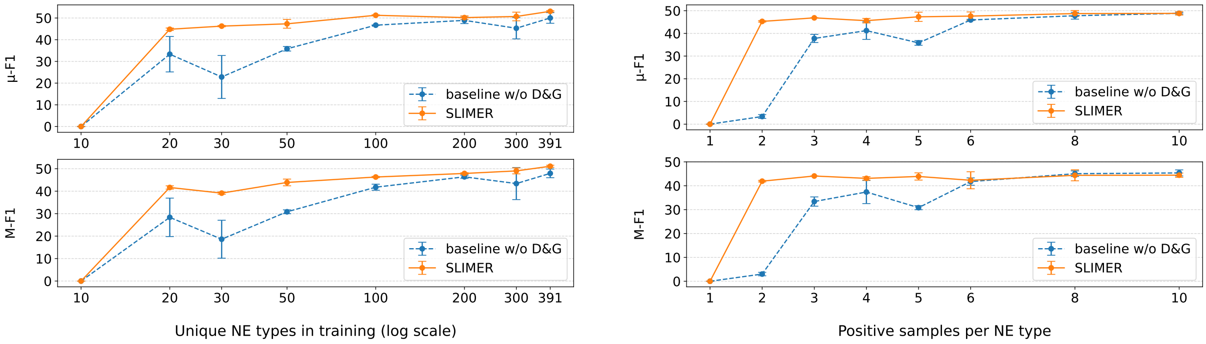

Figure 4: Micro and macro F1 scores of SLIMER and its baseline without D&G on MIT, CrossNER and BUSTER altogether, as we respectively increase the number of unique NEs (left) and the number of samples per NE (right) seen in training.

type for the presence of short, automatically generated definitions of the entity type. These descriptive sentences differ from ours mainly because they are generated within the annotation process and therefore often vary for the same tag. They are also not as structured and detailed as those in SLIMER.

GoLLIE (Sainz et al., 2024) is an LLM based on Code-LLaMA leveraging annotation guidelines formatted in a code-like representation. In the comparison we considered only the 7B version, since it has a number of parameters similar to most of the other approaches, SLIMER included.

GLiNER-L (Zaratiana et al., 2023) is an encoder-only DeBERTa-304M parameters model. We choose their biggest and performing model. Nonetheless, GLiNER-L is by far the smallest one amongst the selected state-of-the-art approaches, yet its quite competitive on OOD zero-shot NER.

GNER (Ding et al., 2024) was released in two versions differing for their backbone LLM: flan-t5-xxl and LLaMA-7B. They are referred as GNER-T5 and GNER-LLaMA, respectively.

4.4 Results

In order to reduce the overlap between training and test classes, we trained SLIMER on the PileNER-type subset described in Section 4.1, which has 391 distinct NEs. As already observed in literature, instruction tuning helps aligning the model to the tasks, and most of the gains in performance are achievable with little amounts of instructions (Wang et al., 2022a; Zhou et al., 2024; Zugarini et al., 2024). Hence, we only picked 5 examples per class from PileNER-type subset. In addition, for each example containing an annotated instance, we also included a negative example, i.e. one without any named entity. Overall, SLIMER was trained on 3910 input sequences, a small fraction of the data typically fed to other state-of-the-art models. To better understand the training set scales of different models, in Figure 5 we depicted them with circles proportional to their training sizes.

Out-of-distribution Domains. Table 3 compares SLIMER against other state-of-the-art LLMs on MIT and CrossNER benchmarks. Differently from models like UniNER, GLiNER and GNER, the overlap between training and test named entity tags is significantly reduced (as shown in Table 2), thus operating in a more fair zero-shot scenario. Despite that, and the fact that we used only a fraction of the training data with respect to other models, SLIMER offers competitive performance, surpassing several of the existing state-of-the-art models. Moreover, our training data is entirely synthetic, whereas models like GoLLIE or UniNER-type+sup also exploit human-annotated labels. The importance of gold-annotated examples can be depicted by observing the 16% absolute increase between UniNER-type and UniNER-type+sup (see Table 3).

Never-seen-before NEs. To experiment the ability of existing models on never-seen-before labels, we extend the zero-shot evaluation on BUSTER, which is characterized by financial entities that are rather far from the more traditional tags observed by all models during training. For simplicity, we limited the evaluation to most performing approaches only, thus we omitted ChatGPT and InstructUIE in this experimentation. Results outlined in Table 4 exhibit an inverse trend with respect to OOD experiments. Indeed, best scoring state-of-the-art models in MIT and and CrossNER, such as the ones from GNER family, both under-perform in BUSTER, also due to their inability to work on

Model	Backbone	#Params	MIT		CrossNER					AVG
			Movie	Restaurant	AI	Literature	Music	Politics	Science	
ChatGPT	gpt-3.5-turbo	-	5.3	32.8	52.4	39.8	66.6	68.5	67.0	47.5
InstructUIE	Flan-T5-xxl	11B	63.0	21.0	49.0	47.2	53.2	48.2	49.3	47.3
UniNER-type	LLaMA-1	7B	42.4	31.7	53.5	59.4	65.0	60.8	61.1	53.4
UniNER-def	LLaMA-1	7B	27.1	27.9	44.5	49.2	55.8	57.5	52.9	45.0
UniNER-type+sup.	LLaMA-1	7B	61.2	35.2	62.9	64.9	70.6	66.9	70.8	61.8
GoLLIE	Code-LLaMA	7B	63.0	43.4	59.1	62.7	67.8	57.2	55.5	58.4
GLiNER-L	DeBERTa-v3	0.3B	57.2	42.9	57.2	64.4	69.6	72.6	62.6	60.9
GNER-T5	Flan-T5-xxl	11B	62.5	51.0	68.2	68.7	81.2	75.1	76.7	69.1
GNER-LLaMA	LLaMA-1	7B	68.6	47.5	63.1	68.2	75.7	69.4	69.9	66.1
SLIMER w/o D&G	LLaMA-2-chat	7B	46.4 ± 1.8	36.3 ± 2.1	49.6 ± 3.2	58.4 ± 1.7	56.8 ± 2.1	57.9 ± 2.1	53.8 ± 1.7	51.3 ± 2.0
SLIMER	**LLaMA-2-chat**	**7B**	$\mathbf{50.9 \pm 0.9}$	$\mathbf{38.2 \pm 0.3}$	$\mathbf{50.1 \pm 2.4}$	$\mathbf{58.7 \pm 0.2}$	$\mathbf{60.0 \pm 0.5}$	$\mathbf{63.9 \pm 1.0}$	$\mathbf{56.3 \pm 0.6}$	$\mathbf{54.0 \pm 0.5}$

Table 3: Comparison of OOD performance for SLIMER and state-of-the-art models on MIT and CrossNER benchmark. With the exception of UniNER-def, all the competitors' results are taken from their respective papers as listed in Section 4.3.

Model	Backbone	#Params	μ-Precision	μ-Recall	μ-F1	M-Precision	M-Recall	M-F1
UniNER-type	LLaMA-1	7B	30.59	40.29	34.78	34.47	45.32	37.58
UniNER-def	LLaMA-1	7B	25.00	51.29	33.62	24.66	48.22	31.80
UniNER-type+sup.	LLaMA-1	7B	31.40	47.53	37.82	32.08	47.52	36.79
GoLLIE †	Code-LLaMA	7B	28.82	26.63	27.68	27.53	22.56	24.13
GLiNER-L	DeBERTa-v3	0.3B	42.55	19.31	26.57	41.16	22.16	24.34
GNER-T5	Flan-T5-xxl	11B	19.31	50.15	27.88	26.64	46.24	30.26
GNER-LLaMA	LLaMA-1	7B	14.68	59.97	23.58	8.87	44.45	14.21
SLIMER w/o D&G	LLaMA-2-chat	7B	44.00 ± 7.84	37.52 ± 3.37	40.41 ± 5.09	39.74 ± 5.82	38.49 ± 2.94	35.90 ± 4.08
SLIMER †	**LLaMA-2-chat**	7B	$\mathbf{47.69 \pm 0.57}$	$\mathbf{43.09 \pm 1.42}$	$\mathbf{45.27 \pm 1.04}$	$\mathbf{42.68 \pm 0.14}$	$\mathbf{41.40 \pm 0.62}$	$\mathbf{40.14 \pm 0.44}$

Table 4: Comparing SLIMER against state-of-the-art models on BUSTER to assess generalization capabilities on never-seen-before NEs. Models leveraging on guidelines are denoted with symbol †.

long input texts [2]. Analogously, GLiNER and GoLLIE struggle in such a benchmark. Notably, GoLLIE's performance is unexpectedly low, given that it is the only other to exploit annotation guidelines and it was fed with the same guidelines we provide to our SLIMER. SLIMER instead, appears to be the most effective in dealing with unseen labels, thanks to its lighter instruction-tuning methodology. Figure 5 clearly delineates such a behaviour.

Increasing the number of training data. We investigated how zero-shot NER performance change as the number of training instances increases. We proceeded in two directions: (1) increasing the number of unique NEs tags, while sampling a fixed total amount of samples per type; (2) varying the number of examples per NE, while keeping fixed the number of distinct categories. In particular, we set 10 examples (5 positive and 5 negative) per class in (1), and we set the number of distinct NEs to 50 in experiment (2). In both cases, we progressively add new classes/examples to the already selected ones. F1 scores are measured by considering MIT, CrossNER and BUSTER as a unique benchmark. By observing the results outlined in

[2]Except for GoLLIE, all the other approaches require input chunking into multiple smaller passages. While the sliding window for most is set to 900 words, for GNER we are limited to work on 150 words per input text.

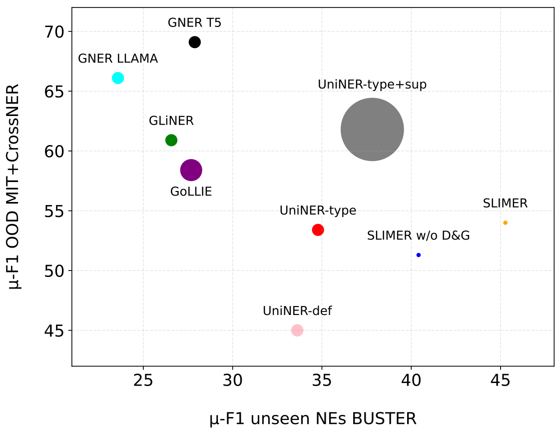

Figure 5: Comparison of state-of-the-art models: F1 scores on unseen NEs in BUSTER (x-axis), OOD evaluation on MIT/CrossNER (y-axis). Circles' size is proportional to the number of examples seen in training by each model.

Figure 4, it emerges that increasing the number of unique entity types somewhat improves results, albeit most of the progresses are achievable with just 20 entity tags. Adding more than two examples per class brings little to no benefits, instead. This confirms what already observed in literature (Wang et al., 2022a; Zhou et al., 2024), further motivating our decision to instruct models on a small portion of the training data.

Dataset	NE	Definition & Guidelines	SLIMER w/o D&G F1	SLIMER-F1	△ F1
BUSTER	Generic Consulting Company	Definition: 'generic consulting company' refers to a business entity that provides non-legal advisory services in areas such as finance, accounting, due diligence, and other professional consulting services., Guidelines: Avoid labeling a company that primarily provides legal services. Exercise caution with company names that include personal names which might be confused with individuals, and consider the context to determine whether the reference is to a company.	2.34	14.78	+12.44
Politics	Person	Definition: 'person' refers to individual human beings, but who are not politicians., Guidelines: In this political context, avoid labelling as 'person' people who are 'politicians'.,	53.87	71.47	+17.60
Movie	Trailer	Definition: 'trailer' refers to a short promotional video that provides a preview or teaser of a forthcoming movie., Guidelines: Label also general entertainment terms like 'preview' or 'teaser'.,	23.44	58.62	+35.18
Science	Chemical Compound	Definition: 'chemical compound' refers to distinct chemical substances composed of two or more elements in fixed proportions., Guidelines: Label entities as 'chemical compound' if they are not proteins or enzymes. Exercise caution with ambiguous terms like 'Almond', which can refer to both a food item and a chemical compound (benzaldehyde). Be aware of complex nomenclature and chemical structures when identifying compounds.	50.32	58.85	+8.53
Restaurant	Amenity	Definition: 'amenity' refers to services, facilities, or features that enhance the convenience, comfort, or enjoyment of a location., Guidelines: When annotating 'amenity', focus on tangible or accessible services and facilities. Avoid labeling abstract concepts, such as 'ambiance' or 'vibe', that are not clearly associated with a specific amenity. Examples of 'amenity' are 'steampunk flavored', 'upscale place' and 'reservation'.	33.38	28.18	-5.20

Table 5: Some examples of definition and guidelines. Absolute F1 gains between SLIMER and its version without definition and guidelines are reported. In red, we highlight performance degradation, in blue positive improvements between 0 and 10 and in green the F1 increases exceeding by over 10 points.

Impact of Definition and Guidelines. As a final analysis, we assess the impact of definition and guidelines in SLIMER. We compare SLIMER with a version of it devoid of definition and guidelines, referred as SLIMER w/o D&G. Results, reported in tables 3,4 and Figure 4 consistently demonstrate how the definition and guidelines are helpful to the model. Indeed, there are improvements in F1 in both OOD and never-seen-before scenarios. Moreover, the absence of guidelines also significantly increase the standard deviation over multiple runs, thus demonstrating that definitions and guidelines also make the learning more consistent and stable. Notably, we can also observe from Figure 4 that in order to reach comparable performance, SLIMER w/o D&G requires more training data. To better understand how definition and guidelines contribute in improving the model, we show some qualitative examples in Table 5[3]. For each example, we report F1 scores obtained by using SLIMER or its version lacking of D&G. Information about an entity type not only helps in detecting novel NEs, like in the case of "Generic Consulting Company", but it can also be beneficial to disambiguate polysemous tags, such as "trailer". Moreover, specifications like the one in "politics", can improve precision, avoiding tagging of instances not strictly belonging to the given class.

5 Conclusions

In this paper, we presented SLIMER, an instruction-tuned LLM for zero-shot NER. With a prompt enriched with definition and annotation guidelines, and a fine-tuning on a restricted set of entity tags, SLIMER, differently from most of the existing models, is specifically designed to better deal with unseen named entity tags. Experiments show that definition and guidelines steer the annotation process, especially on never-seen-before classes, thus yielding better predictions and a more stable learning. Furthermore, SLIMER performs comparably to state-of-the-art approaches, while being trained on a fraction of samples and entity types having little overlap with the test set.

In the future, we plan to broaden the scope of SLIMER to any Information Extraction problem. Moreover, we will investigate solutions to better scale instruction-tuned models for zero-shot NER to large set of entity classes to tag.

[3]Additional examples can be seen in Appendix A.

Limitations

A primary limitation of our approach lies in the instruction-tuning template we have adopted. While extracting the occurrences of a named entity per prompt allows for shorter instructions and a stronger focus on the definitions and guidelines components, it results in the overhead of requiring a number of inference calls per input text equal to the cardinality of the label set. Consequently, our approach does not scale well on datasets with a large number of entity classes.

Another potential limitation could stem from data contamination between the benchmark datasets and the pre-training data of the LLM. However, any performance gap between SLIMER and its baseline can be attributed to the presence of the D&G additional components, as both models share the same pre-training data.

References

Tom Brown et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung et al. 2022. Scaling

instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Bowen Yan, and Min Zhang. 2024. Rethinking negative instances for generative named entity recognition. *Preprint*, arXiv:2402.16602.

Edward J Hu et al. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. A survey on recent advances in named entity recognition. *Preprint*, arXiv:2401.10825.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Alec Radford et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Oscar Sainz et al. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. *Preprint*, arXiv:2310.03668.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Shuhe Wang et al. 2023a. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang et al. 2022b. Super-Natural Instructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei et al. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Junjie Ye et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer. *Preprint*, arXiv:2311.08526.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Andrea Zugarini, Andrew Zamai, Marco Ernandes, and Leonardo Rigutini. 2023. Buster: a “business transaction entity recognition” dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 605–611.

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024. Clue-instruct: Text-based clue generation for educational crossword puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3347–3356, Torino, Italia. ELRA and ICCL.

A Comparing SLIMER to its baseline on some NE classes

In Table 6 we report the F1 scores for some NE classes with the purpose of getting a better insight into the usefulness of the definitions and guidelines for zero-shot NER. We aim to list potential benefits these components can provide and support our thesis with some examples. However, the examples provided are not intended for quantitative assessment, rather they serve as illustrative instances supporting some of our claims.

Different granularity and exceptions. "Every musician is also a person". However, as occurs in CrossNER, there are cases where an individual should be labelled as a person only if it does not fall into the categories of musician, scientist, writer or politician. From Table 6 we can effectively see how the guidelines instruct SLIMER with such requirements and the model improves in both precision and final F1 with respect to its baseline.

Different annotation schemes. Guidelines are key for flexibility to new annotation schemes. As often the case in Zero-Shot NER, a NE in test may require to include or exclude particular instances with respect to what has been trained on; similarly, in supervised-setting, different datasets may adopt different annotation schemes for the same NE (e.g., Zhou et al. (2023) are required to specify in the prompt to which dataset the sample belongs). By carefully formulating the NE definition and guidelines, we can flexibly adapt to the desired behaviour. However, as we can see from the red cases, this can sometimes lead to a drop in performance. We believe that this can be partly due to overly strict guidelines, resulting in higher precision at the expense of lower recall.

Polysemous Named Entities. Guidelines potentially solve the problem of polysemous NE types where, for example, the same NE "title" may denote film titles or nobility titles. We briefly experiment by leaving the NE "title" in the PileNER training set (where it denotes nobility titles) and evaluating on the NE title from MIT-Movie dataset. However, the improvement is only of +2 points, probably because the backbone model is already somehow able to adapt to the correct sense given the context. On the opposite, the polysemous tag "trailer" benefits from the D&G.

Provide external knowledge. Finally, and most importantly in Zero-Shot NER, annotation guidelines may enable the labelling of never-before-seen Named Entities based on the model's ability to adhere to the provided guidelines, thus acting as a source of external knowledge for the model.

Dataset	NE	Definition & Guidelines	baseline-F1	SLIMER-F1	△ F1
BUSTER	Generic Consulting Company	Definition: ’generic consulting company’ refers to a business entity that provides non-legal advisory services in areas such as finance, accounting, due diligence, and other professional consulting services., Guidelines: Avoid labeling a company that primarily provides legal services. Exercise caution with company names that include personal names which might be confused with individuals, and consider the context to determine whether the reference is to a company.	2.34	14.78	+12.44
BUSTER	Selling Company	Definition: ’selling company’ refers to a company that is selling or divesting assets, subsidiaries, or equity to another party as part of a transaction., Guidelines: Be careful when identifying the entity actually doing the selling, as it may not be the main subject of the sentence or document. Pay attention to possessive forms and synonyms such as ’vendor’, ’owner’, or ’parent company’. The company’s role as ’seller’ must be understandable from the sentence in which it appears. Do not label company names which role in the transaction is not evident.	21.13	31.04	+9.91
BUSTER	Buying Company	Definition: ’buying company’ refers to a company that is acquiring another company or its assets through a transaction or merger., Guidelines: When recognizing ’buying company’ entities, focus on the company names directly involved in the acquisition process as buyers, while being careful not to label subsidiaries or companies in other roles. The company’s role as ’buyer’ must be understandable from the sentence in which it appears. Do not label company names which role in the transaction is not evident.	43.09	49.84	+6.75
AI	Person	Definition: ’person’ refers to individuals, including public figures, celebrities, and notable personalities., Guidelines: If a person is working on research (including professor, Ph.D. student, researcher in companies, and etc) avoid labeling it as ’person’ entity.	35.50	41.15	+5.65
AI	University	Definition: ’university’ represents educational institutions that offer higher education and academic research programs., Guidelines: Avoid labeling general concepts such as ’education’ or ’academia’ as ’university’. Exercise caution with ambiguous terms like ’Cambridge’ (can refer to different institutions) and ’Harvard’ (can refer to a person).	64.24	78.38	+14.14
AI	Product	Definition: ’product’ refers to tangible or intangible items, systems, or tools, including but not limited to physical products, software, and industrial machinery, designed for specific functions or applications., Guidelines: Exercise caution when dealing with ambiguous terms like ’Java’ (programming language, island, and coffee). Consider the context to discern the correct entity. Be mindful of generic terms like ’system’ and ’toolkit’ which may require additional context to determine if they fall under the ’product’ category.	20.36	6.02	-14.33
Literature	Event	Definition: ’event’ refers to specific incidents, occurrences, or happenings that take place at a particular time and location, such as festivals, wars, conferences, and award ceremonies., Guidelines: Avoid labeling general or ongoing occurrences, such as ’daily routine’ or ’regular meetings’. Exercise caution with ambiguous terms like ’revolution’ (can refer to a political event or a spinning motion) and ’strike’ (can denote a labor event or a military action).	37.89	54.35	+16.47
Movie	Genre	Definition: ’genre’ refers to a category or classification characterized by specific stylistic, thematic, or content elements., Guidelines: Avoid labeling general terms like ’film’, ’book’, or ’music’ as ’genre’. Focus on labeling genres such as ’science fiction’, ’romance’, ’horror’ and similar. Exercise caution with ambiguous terms that can belong to multiple genres, such as ’drama’ (which can refer to a genre or a situation within a story).,	38.39	46.97	+8.58
Movie	Trailer	Definition: ’trailer’ refers to a short promotional video that provides a preview or teaser of a forthcoming movie., Guidelines: Label also general entertainment terms like ’preview’ or ’teaser’.,	23.44	58.62	+35.18
Movie	Title	Definition: ’title’ refers to names of creative works, such as movies, books, music albums, and artistic productions., Guidelines: Avoid labeling generic words that can be interpreted as common nouns, like ’run’ or ’the’. Exercise caution with potentially ambiguous cases, such as ’Savannah Hilton’ (person with a title name) or ’Apple’ (brand and fruit).	31.42	33.01	+1.58
Politics	Person	Definition: ’person’ refers to individual human beings, but who are not politicians., Guidelines: In this political context, avoid labelling as ’person’ people who are ’politicians’.,	53.87	71.47	+17.60
Music	Award	Definition: ’award’ refers to a recognition or honor given to individuals, groups, or organizations in various fields, such as sports, entertainment, academia, and business., Guidelines: Avoid labeling non-official titles or generic terms like ’best’, ’top’, ’favorite’. Exercise caution with ambiguous terms like ’Oscar’ (could refer to a person or the award) or ’Nobel’ (could refer to the organization or the prize).	67.58	62.27	-5.31
Music	Song	Definition: ’song’ refers to a musical composition with lyrics and melody that can be performed or recorded., Guidelines: Do not label general music-related terms like ’album’ or ’lyrics’. Exercise caution with ambiguous cases like ’Billie Jean’ (which can refer to a person or a song) or ’Guns N’ Roses’ (which can refer to a band or a song).,	59.56	67.39	+7.83
Restaurant	Price	Definition: ’price’ represents the cost or value of a product or service in a given context., Guidelines: Also consider labeling terms like ’cheap’, ’inexpensive’ and similar, when referring to a product or service.,	30.41	42.96	+12.55
Restaurant	Amenity	Definition: ’amenity’ refers to services, facilities, or features that enhance the convenience, comfort, or enjoyment of a location., Guidelines: When annotating ’amenity’, focus on tangible or accessible services and facilities. Avoid labeling abstract concepts, such as ’ambiance’ or ’vibe’, that are not clearly associated with a specific amenity. Examples of ’amenity’ are ’steampunk flavored’, ’upscale place’ and ’reservation’.	33.38	28.18	-5.20
Science	Astronomical Object	Definition: ’astronomical object’ refers to celestial bodies, such as planets, moons, asteroids, and comets, that exist in outer space., Guidelines: Avoid labeling general terms like ’orbit’, ’gravitation’, or ’gravity assist’. Exercise caution with terms that can have multiple meanings, such as ’Moon’ (natural satellite vs. a generic noun) or ’Mars’ (the planet vs. the god of war).	40.18	50.75	+10.57
Science	Chemical Compound	Definition: ’chemical compound’ refers to distinct chemical substances composed of two or more elements in fixed proportions., Guidelines: Label entities as ’chemical compound’ if they are not proteins or enzymes. Exercise caution with ambiguous terms like ’Almond’, which can refer to both a food item and a chemical compound (benzaldehyde). Be aware of complex nomenclature and chemical structures when identifying compounds.	50.32	58.85	+8.53

Table 6: Comparing SLIMER to its baseline w/o D&G on some Named Entities. In red are performance degradations, in blue are positive improvements between 0 and 10, in green are very high improvements over 10 points.