Natural Language Processing Journal 8 (2024) 100099

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

A novel prompting method for few-shot NER via LLMs

Qi Cheng [a], Liqiong Chen [a], Zhixing Hu [b], Juan Tang [a,*], Qiang Xu [a], Binbin Ning [a]

[a] Sichuan Jiuzhou Electronic Group Co., Ltd., Mianyang, Sichuan 621000, China
[b] School of Cyber Science and Engineering, Sichuan University, Chengdu, China

ARTICLE INFO

Keywords:
Large language model
Named entity recognition
Natural language processing
Prompt method
Deep learning

ABSTRACT

In various natural language processing tasks, significant strides have been made by Large Language Models (LLMs). Researchers leverage prompt method to conduct LLMs in accomplishing specific tasks under few-shot conditions. However, the prevalent use of LLMs' prompt methods mainly focuses on guiding generative tasks, and employing existing prompts may result in poor performance in Named Entity Recognition (NER) tasks. To tackle this challenge, we propose a novel prompting method for few-shot NER. By enhancing existing prompt methods, we devise a standardized prompts tailored for the utilization of LLMs in NER tasks. Specifically, we structure the prompts into three components: task definition, few-shot demonstration, and output format. The task definition conducts LLMs in performing NER tasks, few-shot demonstration assists LLMs in understanding NER task objectives through specific output demonstration, and output format restricts LLMs' output to prevent the generation of unnecessary results. The content of these components has been specifically tailored for NER tasks. Moreover, for the few-shot demonstration within the prompts, we propose a selection strategy that utilizes feedback from LLMs' outputs to identify more suitable few-shot demonstration as prompts. Additionally, to enhance entity recognition performance, we enrich the prompts by summarizing error examples from the output process of LLMs and integrating them as additional prompts.

1. Introduction

Large Language Models (LLMs), due to their extensive training scale and numerous model parameters, exhibit significant effectiveness across a variety of generative tasks, including machine translation (Zhang et al., 2023a) and human–computer dialogue (Vemprala et al., 2023). Meanwhile, LLMs can also be employed to tackle intricate challenges within specialized domains, including clinical medical assistance (Jeblick et al., 2023), climate scenario generation (Biswas, 2023), and investment analysis (Leippold, 2023), etc. As LLMs can rapidly execute specified downstream tasks in zero-shot or few-shot scenarios via in-context learning(ICL), there has been relentless exploration of LLMs' potential in other natural language processing tasks in recent years, such as Named Entity Recognition.

Named Entity Recognition (NER), an important branch of information extraction, has wide-ranging applications in numerous practical intelligent systems. Currently, NER methods that are based on deep learning primarily use richly annotated datasets and employ supervised learning to obtain satisfactory recognition performance. However, these methods tend to underperform when dealing with entities that are not included in the annotated datasets. Researchers have proposed employing pre-trained models or contrastive learning techniques to tackle few-shot and cross-domain NER tasks. Although this approach can offer some savings in training costs and computational resources, it still necessitates model fine-tuning or retraining for varying datasets, and the achieved recognition performance remains suboptimal. Given the practical application scenarios, we aim for models that can adjust to different entity categories with minimal sample data and training costs. This issue has ushered in new possibilities for exploration with the advent of Large Language Models.

Conversely, the core architecture of LLMs is a generative model rooted in the Transformer structure, which inherently excels at handling generative tasks. In LLMs, prompts, which are constructed based on the in-context learning method, play a pivotal role in instructing the models. This method primarily feeds LLMs with few-shot demonstrations (input–output pairs), and instructs LLMs to perform specific tasks predicated on few-shot instances. However, due to the training objectives of the NER task were inconsistent with those of the generation task, resulting in the direct demonstration of the existing prompts may not yield the anticipated results in NER tasks. Moreover, as the demonstration of few-shot demonstrations in LLMs can be viewed as a fine-tuning process for the model, the selection of demonstrations is of utmost importance.

After conducting the aforementioned analysis, we propose a novel prompt method with LLMs to address NER tasks in few-shot scenarios.

* Corresponding author.
E-mail addresses: 2022340312cq@my.swjtu.edu.cn (Q. Cheng), qianyuhuan910@sina.com (L. Chen), hu1603324796@gmail.com (Z. Hu), cynthia.landseer@gmail.com (J. Tang), DarkBlueHSFS@outlook.com (Q. Xu), 491349717@qq.com (B. Ning).

https://doi.org/10.1016/j.nlp.2024.100099
Received 20 June 2024; Received in revised form 4 August 2024; Accepted 20 August 2024

For NER tasks, we standardized the existing prompts and proposed a structured prompt containing task definition, few-shot demonstration, and output format, and optimized the prompts specifically. In addition, we propose a selection approach for few-shot demonstrations. This approach utilizes LLMs to filter datasets for NER and integrates NER task metrics. Ultimately, it selects demonstrations with high accuracy and better alignment with LLMs' encoding representation. Additionally, we propose augmenting existing prompts based on errors encountered during practical usage of LLMs, including extracting entity content beyond the input sentence and misclassifying or misidentifying entity categories.

In general, the contributions of this paper can be summarized as follows:

- Proposed a standardized prompt method and customized the content of the prompt for NER tasks.
- Introduced a method for selecting few-shot demonstrations in prompts.
- Designed additional prompts specifically targeting errors in LLMs for NER tasks.

2. Related work

2.1. Named Entity Recognition

Traditional Named Entity Recognition (NER) methods can be roughly categorized into tagging-based, span-based, and generative-based approaches. Tagging-based methods treat NER as a sequence labeling task, predicting corresponding labels for each token in a sentence. In these methods, researchers have proposed various structures to address both label accuracy and nested sequence issues. Ju et al. (2018) proposed a dynamic stacked tagging layer to identify nested entities. Wang et al. (2020) employed a bidirectional pyramid structure to construct a tagging layer for entities, enabling more precise identification of nested relationships among entities. Span-based methods, on the other hand, regard NER as a span-oriented classification task. Some researchers (Sohrab and Miwa, 2018) enumerate and extract spans from input sentences, then classify these spans to identify different entities within the sentence. Others utilize boundary identification (Zheng et al., 2019; Tan et al., 2020) to extract spans. Shen et al. (2021) proposed a two-stage entity identifier, which filters existing spans through boundary regression and utilizes an entity classifier to label corresponding categories on the filtered spans. Generative-based methods (Yan et al., 2021; Lu et al., 2022; Zhang et al., 2022) treat NER as a generative task, where entities to be extracted are considered as the predicted content through a seq2seq model structure. This approach allows for the extraction of flat and nested entities within a unified framework.

Furthermore, researchers are exploring methods based on few-shot approaches. ProtoBERT (Snell et al., 2017) developed a prototype network to categorize representation distances across various classes., while NNShot (Yang and Katiyar, 2020) accomplishes entity sequence labeling by duplicating labels from neighboring entities. Expanding on these studies, recently, CONTAINER (Das et al., 2022) and COP-NER (Huang et al., 2022) refined token representations through contrastive learning, thereby bolstering entity recognition accuracy in few-shot settings. Researchers have expanded the application scenarios of few-shot learning by delving deeper into entity recognition within cross-domain settings. CPNER (Chen et al., 2023), which uses collaborative prefix tuning to learn domain-specific prefixes that can be swapped flexibly to perform NER; FactMix (Yang et al., 2022), which uses a model-agnostic data augmentation strategy to improve generalization; LANER (Hu et al., 2022), which uses a dual attention module that incorporates labeling features, improving label transferability significantly.

In traditional methods, researchers address issues such as entity recognition accuracy and nested entities from various perspectives. Nonetheless, attaining commendable recognition performance using these methods typically demands a significant amount of annotated data for training. While existing few-shot techniques can economize to some extent in terms of training costs and computational resources, they still require fine-tuning or retraining the model. Moreover, the actual recognition performance of these methods also needs improvement. The integration of LLMs introduces a novel learning approach where LLMs are directly applied to different tasks through prompt statements, brings new exploration possibilities in NER within few-shot scenarios.

2.2. Large language model prompt method

While large language models (LLMs) can be fine-tuned similarly to pre-trained models to conduct various downstream tasks. Given the substantial training resources demanded by LLMs, retraining LLMs may not present the most optimal application solution. An alternative strategy involves formulating prompt statements using the in-context learning (ICL) method, enabling the exploitation of LLMs without the need for retraining. Initially, researchers like Brown et al. (2020) and Min et al. (2022) explored the potential of LLMs in handling diverse natural language tasks within few-shot exsamples. Subsequently, Wei et al. (2022) introduced the Chain-of-Thought (CoT) prompting method to enrich the logical reasoning abilities of LLMs, with further extensions of the CoT method to multimodal tasks by Zhang et al. (2023b). In recent studies, Diao et al. (2023) have shown that improved prompts and task-specific examples can significantly boost the performance of LLMs.

The ICL method within large language models enables researchers to utilize LLMs across diverse natural language processing tasks without model retraining. Nevertheless, current approaches, predicated on a single prompt format (like few-shot demonstrations or CoT prompts), fall short of achieving entity recognition performance comparable to existing few-shot NER methods. Hence, for NER, it is crucial to develop more standardized and precise prompts to fully leverage the capabilities of LLMs.

3. Method

3.1. NER task description and model overview

In Named Entity Recognition (NER), the model is required to identifying and outputting entities along with their respective categories within a provided sentence x. These entities are then extracted into a structured output format $y = \{(e_0, t_0), (e_1, t_1), \dots\}$, where e represents the entity span and t denotes the defined entity category in the dataset. Only by simultaneously extracting entities and accurately classifying them can the NER tasks be considered successfully completed.

In LLMs, output generation is usually based on a conditional probability model, with the formula $p(y|x, c)$, where y represents the output, x represents the input, and c represents the prompt. Given the input text (prompt), the model calculates the conditional probability distribution of the output text (generated result) and selects the most probable output. An effective prompting method can guide the model to focus on key information, reduce ambiguity, provide more targeted instructions, thereby influencing the process of conditional probability calculation, and increasing the likelihood of generating more coherent and accurate outputs.

The model's overall structure, as shown in Fig. 1, comprises three modules: initial prompt construction, demonstration selection, and additional prompt. During the initial prompt construction phase, the prompts is divided into three standardized components: task definition, few-shot demonstration, and output format. Thereafter, the training dataset, along with the prompts, is fed into the LLMs for entity recognition. Subsequently, based on the identification results of the training set, construct demonstrations for selection and additional prompts. The

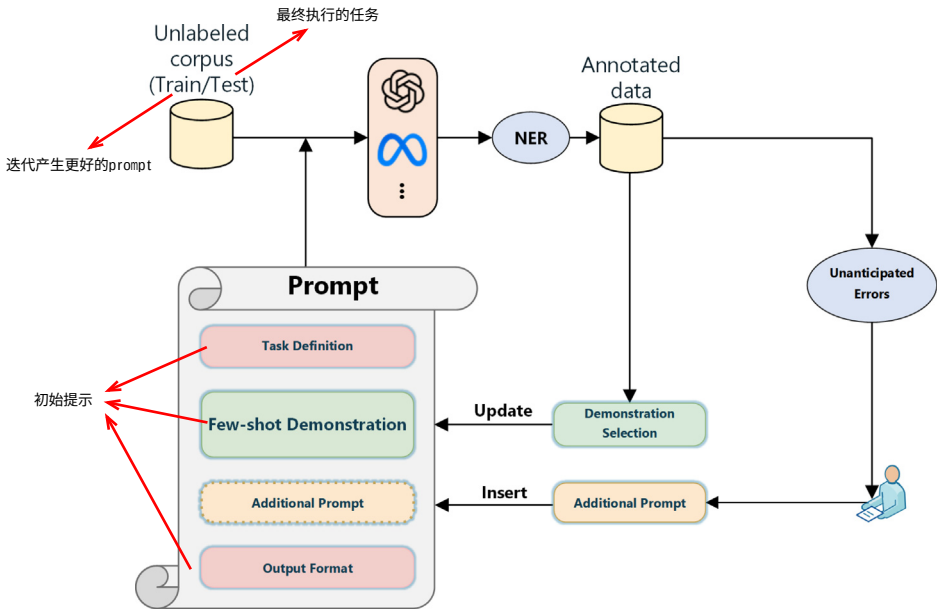

Fig. 1. The overview of proposed model. **Unanticipated Errors** is artificially summarizing errors in extraction scenarios. **Update** is replacing the existing demonstration in the prompt with the selected one. **Insert** is adding additional prompt to prompts to resolve unanticipated errors.

content from demonstration selection and additional prompts is then integrated back into the initial prompts to update the prompts. The comprehensive prompt is shown in Fig. 2. Finally, the comprehensive prompt is employed with the test dataset to execute entity recognition using LLMs.

3.2. Initial prompt construction

3.2.1. Task definition

In task definition, to conduct the LLMs for NER tasks, we have adopted a role-playing paradigm (Shanahan et al., 2023) for the design of the first prompt sentence, which is as follows:

> *You are now an entity recognition model. Always answer as helpfully as possible.*

The subsequent part of the Task definition is designed to *Entity Category* and *Category Description*, which is as follows:

> *"Entity Category": Category Description*
>
> *"Entity Category": Category Description*
>
> ...

Entity Category denotes the defined types of entities requiring identification. *Category Description* providing comprehensive descriptions of these categories. For instance: *"Company": All company names are entities and a company is a business entity that makes money by selling goods and services*. Moreover, due to input length restrictions in some LLMs (e.g., GPT-3 Brown et al., 2020 and Llama2 Touvron et al., 2023 limited to 4096 tokens), simplification may be necessary if subsequent prompts are excessively lengthy. The simplified category definition is as follows:

> *Definition An entity is a "Entity Category", "Entity Category", "Entity Category", ...*

In the simplified definitions, the category meanings were intentionally omitted. This can be attributed to the substantial knowledge base inherent in LLMs, which enables acquisition of the definitions corresponding to the categories, even when only the entity categories are provided. Consequently, enabling LLMs to recognize entities. To adapt this prompt in different domains simply requires adjustments to this definition, reducing computational expenses under fine-tuning strategies and enhancing the method's flexibility.

3.2.2. Few-shot demonstration

Within in-context learning, the selection and presentation of specific demonstrations is crucial. Employing few-shot demonstration typically produces better outcomes compared to zero-shot instances. This is because when provided with few-shot demonstration, it is often necessary to provide both the input and the output. Upon receiving such examples, LLMs have the capability to directly obtain the precise predictive results required for predetermined tasks, and carry out memorization and inference based on this foundation. Hence, for NER, specific configurations are essential when introducing few-shot demonstration. The few-shot demonstration is as follows:

> *Example: [Input Sentence]*
>
> *Output: Entity | Entity Category | True/False [Whether it is the target entity] | Explanation*

In the demonstration's output, we specify that it must encompass both the entity's content and its corresponding category. Furthermore, LLMs are required to self-evaluate their output, accompanied by the provision of pertinent reasons. The conceptualization behind this stems from the Chain-of-Thought prompting method (Wei et al., 2022). This approach will also be adapted in Section 3.2.3 to standardize the format of the output.

3.2.3. Output format

For LLMs, constraining output is essential. As open domain dialogue models, LLMs may produce unrelated answers without a defined output format. While few-shot examples can enable LLMs to learn the output format within the demonstration, without prompts to constrain the output format, LLMs may still produce unrelated answers. Therefore, we specify that each line in the output format should indicate to a candidate entityand its corresponding entity type. The prompt construction of entity output format is as follows:

> *Given the paragraph below, identify the possible entities and their categories based on my instructions, list each entity in the format "Talbott | True | as it is a name of a person (PER).and so on".*

Furthermore, based on previous research on Chain-of-Thought prompt method, intentionally conducting the model to engage in reasoning and explain the rationale behind its conclusions enhances the model's performance in subsequent tasks. Within the permissible length constraints, we additionally require LLMs to offer an explanation of why it chose a particular entity. The construction of this prompt is as follows:

可以尽量省略对实体的定义描述

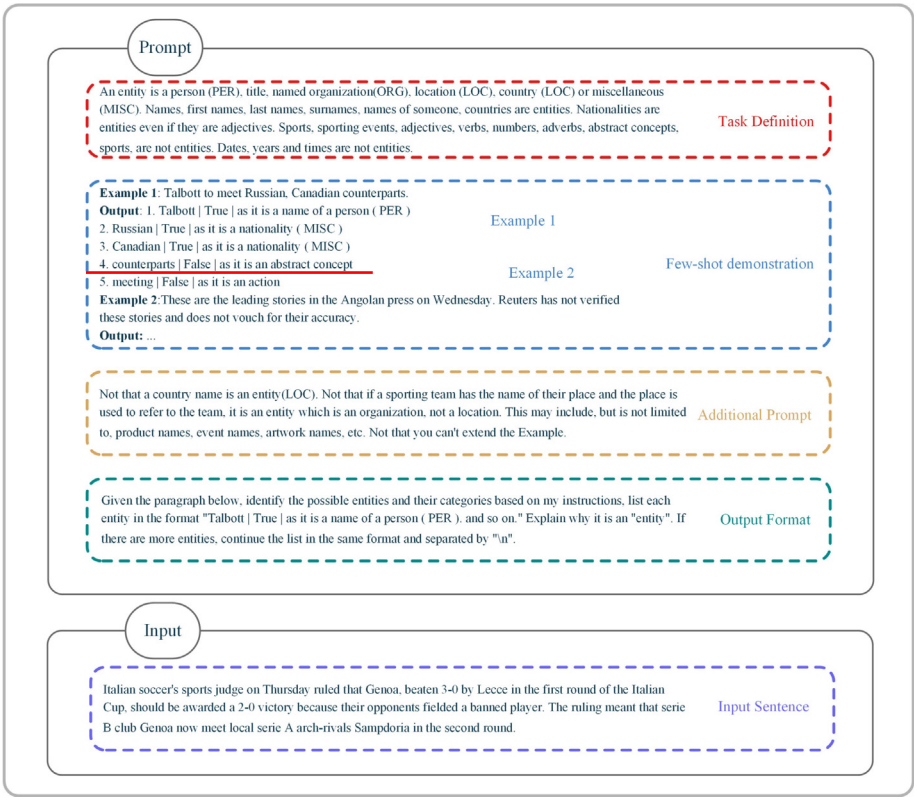

Fig. 2. The example of prompt. The prompt consists of four components: (1) **Task Description** is surrounded by a red rectangle. (2) **Few-shot Demonstrations** is surrounded by a yellow rectangle. (3) **Additional Prompt** is surrounded by a blue rectangle. (4) **Output Format** is surrounded by a green rectangle.

Explain why it is an "entity". If there are more entities, continue the list in the same format and separated by "\n".

3.3. Demonstration selection

As mentioned in Section 3.2.2, few-shot demonstrations have been shown to significantly enhance the performance of LLMs. Hence, it is important to select the most suitable demonstrations from vast datasets. Traditional methods involve random selection from the dataset or employing K Nearest Neighbors(KNN) to choose demonstrations with the most similar sentence or entity representations. Under the massive pre-training scale, demonstrations selected using KNN may not be the optimal selection within the well-trained representations of LLMs. Additionally, as the usage process of LLMs is similar to black-box testing, the effectiveness of entity extraction can only be determined through input–output analysis. To LLMs, we propose a simple and convenient strategy for filtering few-shot demonstrations. In the first step, we use the initial prompt construct by Section 3.2 to predict the output of the complete training set, which illustrated in Eq. (1), where e is the entity span, t is the entity category predicted by LLMs, n is the total number of entities identified by LLMs, $LLMs$ is using LLMs for NER tasks, D_{train} is the training set and $Prompt$ is the initial prompt which contains task definition, few-shot demonstration, additional prompt and output format. The number of few-shot demonstration can be one or more.

$$y = \{(e_0, t_0), (e_1, t_1), \ldots, (e_n, t_n)\} = LLMs(D_{train}, Prompt) \quad (1)$$

Next, we calculate the F1 score based on the LLMs' predicted results and extract the top-k demonstrations from the training set based on these F1 score predictions, which as shown in Eq. (2), where y_{filter} is extracted results from y, F_1 is F1 score evaluation metrics, Top_k is k maximum results of F1 score and we set k to 1~2 or 5~10 to adapt the experimental setup in Section 4.2.

$$y_{filter} = Top_k(F_1(y)) = \{(e_0, t_0), \ldots, (e_k, t_k)\}, y_{filter} \subseteq y, k < n \quad (2)$$

Throughout the filtering process, we exploit the black-box testing nature of LLMs. For LLMs under in-context learning, the distinction between training and testing sets is absent. As the input text content remains uninvolved in the model's training or fine-tuning, all data can be treated as part of the testing set. The purpose of the first step is to determine which demonstrations from the training set yield superior extraction outcomes align with LLM encoding representation. The subsequent step incorporates the utilization of evaluation metrics to further filter for demonstrations with better extraction results. Through steps like these, we can obtain few-shot demonstrations that benefit the LLM encoding representation and have high F1 scores.

不会存在数据泄露

3.4. Additional prompt

Although after standardizing the prompts of the LLMs, we have achieved quite satisfactory recognition results in NER tasks. Unanticipated errors may arise during practical extraction scenarios, these errors include: (1) extracting entity content beyond the input sentence; (2) misclassifying or misidentifying entity categories. To address such misjudgments, we propose to enhance the original prompts with additional content that present in advance potential error scenarios to LLMs. The example of additional prompts can be found in Fig. 2. Each additional prompt is tailored to address one of the aforementioned errors. To address the first error type, the additional prompt is crafted as follows:

The entity you provide must be a word that exists in the "paragraph"

The main purpose of this prompt is to restrict LLMs to extract entities within the specified input context, making it applicable to general NER tasks. Regarding the second type of error, the prompt template is formulated as follows:

[Entity Scope/Entity Description] are likely to be an entity

[Entity Scope/Entity Description] are not entities

The content above should be customized according to the different categories and dataset, replacing *[Entity Scope/Entity Description]* with the incorrect entity extraction results. For instance, concerning location entities, the content could be *"A proper noun with capital letters are likely to be an entity"*. In datasets with person names, location names, etc., the content could be *"Dates, years and times are not entities"*.

4. Experiment

In all experimental sections, we utilized the GPT-3.5 (Brown et al., 2020) model, particularly the text-davinci-003 variant, along with the GPT-4 (Achiam et al., 2023) and Llama2-13B (Touvron et al., 2023) models. The evaluation metric employed to assess experimental results is the widely used **Micro-F1** score in NER. The few-shot NER selection of baselines and comparison methods varied depending on the specific task due to the testing of task performance across diverse scenarios.

4.1. Experiment datasets

To assess the efficacy of the proposed NER method, we selected three distinct datasets: CoNLL-2003, Few-NERD, and CrossNER. These datasets facilitate the comprehensive evaluation of our method's performance across various NER task scenarios.

CoNLL-2003 (Sang and De Meulder, 2003), a standard benchmark dataset for NER, encompasses four defined entity categories: Location, Organization, Person, and Miscellaneous. The dataset statistics are as follows: the training set comprises 14,987 sentences, the development set consists of 3,466 sentences, and the test set contains 3,684 sentences.

Few-NERD (Ding et al., 2021) is a dataset specifically crafted for NER, featuring an extensive annotation of instances (Wei et al., 2022). It encompasses eight defined entity categories: People, Art, Product, Location, Event, Building, Organization, and Miscellaneous. The Few-NERD(INTRA) statistics are as follows: the training set comprises 99,519 sentences, the development set consists of 19,358 sentences, and the test set contains 44,059 sentences.

CrossNER (Liu et al., 2021) is a cross-domain NER dataset, a fully-labeled collection of NER data spanning over five diverse domains (Politics, Natural Science, Music, Literature, and Artificial Intelligence) with specialized entity categories for different domains. The dataset statistics are as follows: the training set comprises 700 sentences, the development set consists of 2,121 sentences, and the test set contains 2,497 sentences.

4.2. Results in few-shot NER

For few-shot NER, we compare our proposed method with a series of traditional supervised learning baseline methods, including ProtoBERT, NNShot, StructShot, CONTAINER and COPNER. Additionally, we have conducted experiments examining the impact of general-purpose prompt.[1] (with few-shot demonstration Brown et al., 2020) on LLMs. These approaches represent prevalent baseline models and recent research proposals frequently employed for comparison in few-shot scenarios. The experimental findings are detailed in Table 1.

Table 1
Few Shot results on the CoNLL and the FewNERD dataset on the INTRA 10-way task. Results show micro-F1 scores. We show the **best** for each column in bold.

models	CoNLL		Few-NERD(INTRA)	
	1~2 shot	5~10 shot	1~2 shot	5~10 shot
ProtoBERT (Snell et al., 2017)	49.9	61.3	19.76	34.61
NNShot (Yang and Katiyar, 2020)	61.2	74.1	21.88	27.67
StructShot (Yang and Katiyar, 2020)	62.4	74.8	25.38	26.39
CONTAINER (Das et al., 2022)	61.2	75.8	33.84	47.49
COPNER (Huang et al., 2022)	67.0	74.9	44.13	51.55
few-shot(Lama2-13B) (Brown et al., 2020)	32.4	34.5	11.7	15.8
few-shot(GPT-3.5) (Brown et al., 2020)	49.7	53.2	16.1	18.9
few-shot(GPT-4) (Brown et al., 2020)	66.7	69.3	21.9	26.3
ours(Lama2-13B)	58.4	61.9	37.4	44.2
ours(GPT-3.5)	78.4	81.7	64.41	67.1
ours(GPT-4)	**81.3**	**88.7**	**74.62**	**78.3**

数据集

In low-resource scenarios with 1–2 shot settings, our proposed method demonstrated a 21.3% improvement[2] in F1 score compared to other comparison methods. This can be attributed to the assistance derived from pre-training on large language model knowledge bases, which simplifies the recognition of target entities for greater convenience. Conversely, other methods necessitate more training data to optimize the models. Consequently, with an increase in training data to 5–10 samples, all comparison methods exhibited varying degrees of enhancement. With the support of an increased number of demonstrations, the prompts incorporated in LLMs can further facilitate LLMs in developing a cognitive comprehension of entities, consequently augmenting the precision of entity recognition. Furthermore, using LLMs of various model sizes for entity recognition under the same prompt method showed different results. Specifically, the implementation of GPT-3.5 and GPT-4 resulted in an improvement of at least 41.2% and 88.3% in Conll dataset and Few-Nerd dataset compared to the usage of Llam2-13B[3]. This finding suggests that larger-scale models in LLMs generally possess a significant edge in achieving accurate entity recognition in NER.

4.3. Results in cross domain NER

For cross domain NER, we chose FactMix, LANER and CPNER as comparison methods, which are recent proposals specifically designed for cross domain NER. Additionally, we have conducted experiments examining the impact of general-purpose prompt(with few-shot demonstration Brown et al., 2020) on LLMs. Regarding the datasets, we followed the approach proposed by Liu et al. (2021), employing the subset dataset CrossNER, derived from the CoNLL dataset, tailored for cross domain NER. Detailed experimental results are available in Table 2.

Based on the results, our proposed approach showcased an enhancement in F1 score contrasted to the comparison methods in the Politics, Literature, and Music domains. In the AI and Sciences domains, our method achieved comparable performance with established comparison methods. We posit that fields such as AI and Sciences, compared to areas like politics and music, inherently contain more technical terms. This complexity could potentially lead to a relative reduction in the performance of large language models in entity extraction, irrespective

[1] Since the focus of this paper is primarily on few-shot NER and to streamline the statistical analysis of model predictions, we structured the prompts used in this paper based on general-purpose few-shot prompt and also used this prompt in Section 4.3 Which is shown as follows: Example 1: Talbott to meet Russian, Canadian counterparts. Output: 1. Talbott | True | as it is a name of a person (PER) 2. Russian | True | as it is a nationality (MISC) 3. Canadian | True | as it is a nationality (MISC) 4. counterparts | False | as it is an abstract concept 5. meeting | False | as it is an action

[2] On the CoNLL dataset under 1–2 shot, 21.3% is obtained by subtracting the F1 score of COPNER from the F1 score of ours (GPT-4) and then dividing by the F1 score of COPNER.

[3] On the CoNLL dataset, under 1–2 shot and 5–10 shot conditions, the average improvement of ours (GPT-4) over ours (Llama2-13B) is calculated, as well as the average improvement on the Few-NERD(INTRA) dataset under 1–2 shot and 5–10 shot conditions.

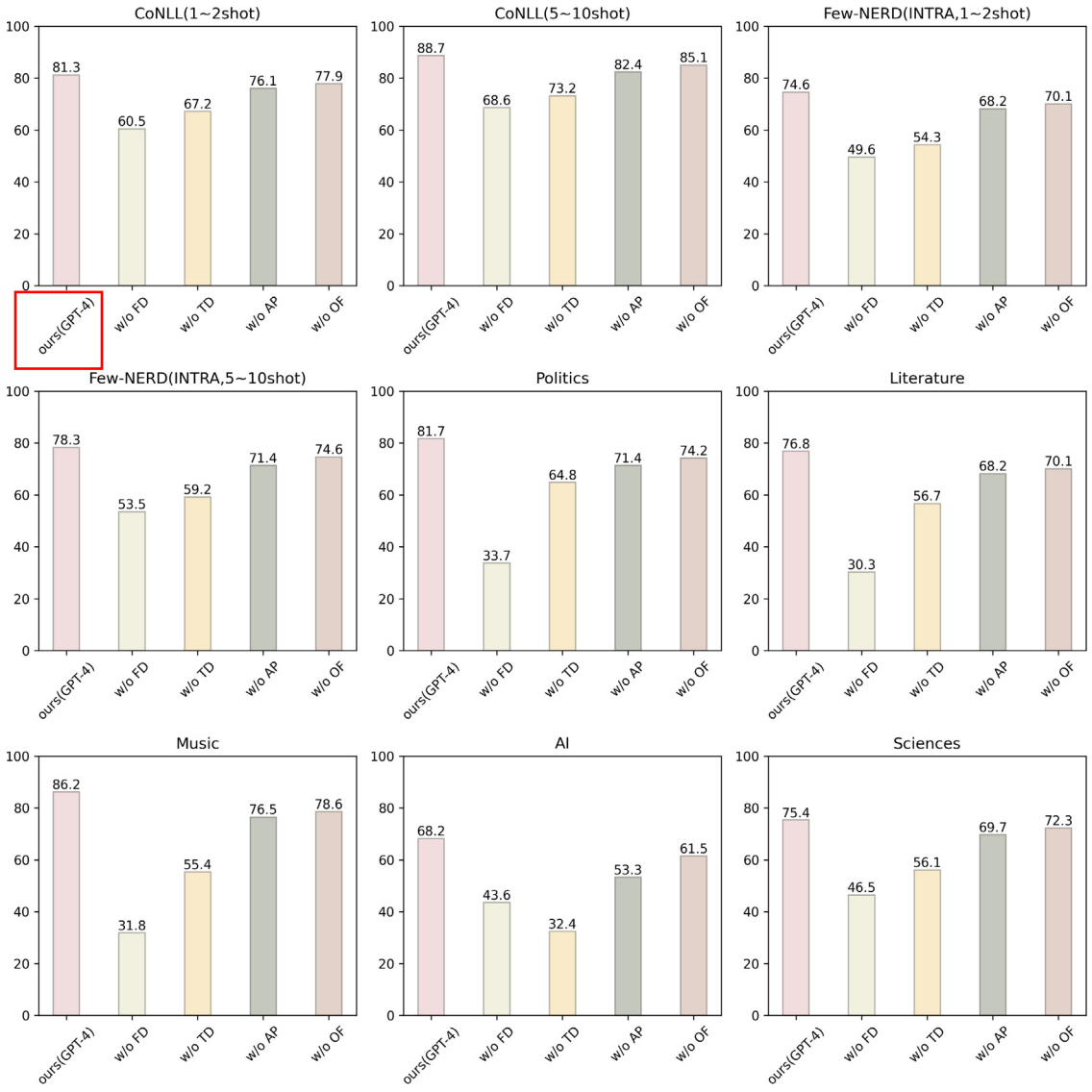

Fig. 3. Ablation of different components in prompt construction on GPT-4, and few-shot demonstrations were uniformly selected following the methodology detailed in Section 3.2.

Table 2
Cross Domain results on CrossNER dataset. Results show micro-F1 scores. We show the **best** for each column in bold.

Method	Politics	Literature	Music	AI	Sciences
FactMix (Yang et al., 2022)	44.66	76.4	23.75	32.09	34.13
LANER (Hu et al., 2022)	74.06	71.11	78.78	65.79	71.83
CPNER (Chen et al., 2023)	76.35	72.17	80.28	66.39	76.83
few-shot (Llama2-13B)	29.2	28.6	30.1	20.3	22.5
few-shot (GPT-3.5)	42.5	40.3	48.6	24.5	35.9
few-shot (GPT-4)	58.3	46.4	54.7	27.7	36.2
ours (Llama2-13B)	44.5	45.9	45.8	38.6	51.2
ours (GPT-3.5)	73.3	67.4	79.5	62.8	69.3
ours (GPT-4)	**81.7**	**76.8**	**86.2**	**68.2**	**75.4**

of their training scale. Yet, due to the absence of public access to the pre-training data and parameters of these large models, deduction of concrete reasons for this observed phenomenon remains elusive. From the results, the recognition effect of LLMs is not much different from the existing methods, so we believe this result is also acceptable. In summary, performing entity recognition across various domains demonstrates the capacity of LLMs to effectively achieve recognition in cross-domain tasks.

4.4. Ablation study

The proposed method in this paper introduces two key enhancements to existing LLMs prompting techniques. Firstly, it standardizes the formulation of NER prompt statements and incorporates additional prompts to assist LLMs in mitigating errors during recognition. Secondly, it optimizes the process of demonstration selection within current prompts. Consequently, ablation experiments were conducted to assess the individual contributions of each prompt construction component and various demonstration selection methods.

4.4.1. Impact of different components in prompt construction

The prompts constructed in this paper encompass four components: task definition, few-shot demonstration, additional prompt, and output format. The ablation experiments investigated the influence of these prompt components on NER tasks by selectively removing distinct components within the prompts in Fig. 2. Including removing few-shot demonstration(w/o FD), removing task definition(w/o TD), removing additional prompt(w/o AP), removing output format(w/o OF), and the results illustrated in Fig. 3.

The experimental results indicate that each component of the prompt construction method proposed in this study contributes to different

Table 3
Demonstration Selection Ablation on Few-shot dataset on GPT-4.

models	CoNLL		Few-NERD(INTRA)	
	1~2 shot	5~10 shot	1~2 shot	5~10 shot
ours (GPT-4)	**81.3**	**88.7**	**74.62**	**78.3**
w/ random selection	71.45	73.49	65.7	69.3
w/ KNN	78.89	84.64	70.1	73.8

Table 4
Demonstration Selection Ablation on CrossNER dataset on GPT-4.

Method	Politics	Literature	Music	AI	Sciences
ours (GPT-4)	**81.7**	**76.8**	**86.2**	**68.2**	**75.4**
w/ random selection	70.6	68.5	75.3	54.9	68.7
w/ KNN	76.4	72.1	81.7	62.3	73.5

extents of enhancement in NER tasks. Removing either the task definition or few-shot demonstration leads to a significant decline in the recognition effectiveness of LLMs. During the testing process, we discovered that the absence of specific definitions for entity types increases the likelihood of LLMs misclassifying entity categories. Moreover, the absence of few-shot demonstration results in the model erroneously labeling non-entity sentence content. Thus, the removal of either the task definition or few-shot demonstration within the prompts considerable impact the performance of LLMs in NER tasks.

Building on the simultaneous use of task definition and few-shot demonstration, the incorporation of additional prompts and output formats can further augment LLMs' performance in NER tasks. However, there is also a disparity in the degree of enhancement between these two components. Specifically, additional prompt are customized based on specific errors detected by LLMs during testing across diverse datasets. However, the output format primarily imposes uniform constraints on all datasets, without additional conditions tailored to specific domains or data. Therefore, the enhancement effect of additional prompt should be better than the output format, and the experimental results have also validated our inference.

4.4.2. Impact of demonstration selection methods

To assess the efficacy of the demonstration selection method proposed in this paper, we performed experiments using various selection schemes, include random selection, selection based on KNN sentence representations, and our proposed method. The experimental results are illustrated in Tables 3 and 4.

The experimental results demonstrate that the demonstration selection approach introduced in this paper outperforms two conventional methods on different datasets. While random selection is straightforward and uncomplicated, the shortcoming is obvious: there is no guarantee that selected demonstration are semantically close to the input. Conversely, the KNN method enhances the correlation among sentence representations, ensuring a level of relevance between the chosen demonstration and the input sentence. Nevertheless, as outlined in Section 3.3, demonstration derived from KNN may not always be the optimal choice for fitting LLMs training representation and may not yield optimal accuracy in entity recognition. Hence, the method proposed in this paper carefully considers both aspects when selecting demonstration from the dataset, filtering out demonstration that favorable to LLMs' encoding capabilities and exhibit high F1 scores.

5. Conclusion

In this work, we proposed a novel prompting method based on LLMs, successfully applying it to few-shot NER tasks. We optimize the prompts for LLMs and propose a standardized prompt structure tailored specifically for NER tasks. We customize the content of each prompt component while optimizing the strategy for selecting few-shot demonstrations. Moreover, we address errors in prompts during actual testing by proposing additional prompts to enhance LLMs recognition performance. We validate the effectiveness of our proposed method through extensive experiments, which demonstrate its ability to improve LLM recognition performance in few-shot NER tasks. In future research, we will explore better prompts and further analyze errors that occur, making them more widely applicable in NER tasks. Additionally, we plan to investigate integrating LLMs with existing NER methods to tackle more complex entity recognition tasks, including nested entities.

CRediT authorship contribution statement

Qi Cheng: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Liqiong Chen:** Validation, Investigation, Data curation. **Zhixing Hu:** Writing – review & editing, Visualization, Validation, Software, Investigation, Data curation. **Juan Tang:** Supervision, Resources, Project administration, Funding acquisition. **Qiang Xu:** Writing – review & editing, Data curation. **Binbin Ning:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Biswas, S.S., 2023. Potential use of chat gpt in global warming. Ann. Biomed. Eng. 51 (6), 1126–1127.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 1877–1901.

Chen, X., Li, L., Qiao, S., Zhang, N., Tan, C., Jiang, Y., Huang, F., Chen, H., 2023. One model for all domains: collaborative domain-prefix tuning for cross-domain NER. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 5030–5038.

Das, S.S.S., Katiyar, A., Passonneau, R.J., Zhang, R., 2022. Container: Few-shot named entity recognition via contrastive learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6338–6353.

Diao, S., Wang, P., Lin, Y., Zhang, T., 2023. Active prompting with chain-of-thought for large language models. arXiv preprint arXiv:2302.12246.

Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., Zheng, H., Liu, Z., 2021. Few-NERD: A few-shot named entity recognition dataset. 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). pp. 3198–3213.

Hu, J., Zhao, H., Guo, D., Wan, X., Chang, T.-H., 2022. A label-aware autoregressive framework for cross-domain NER. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 2222–2232.

Huang, Y., He, K., Wang, Y., Zhang, X., Gong, T., Mao, R., Li, C., 2022. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 2515–2527.

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A.T., Topalis, J., Weber, T., Wesp, P., Sabel, B.O., Ricke, J., et al., 2023. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur. J. Radiol. 1–9.

Ju, M., Miwa, M., Ananiadou, S., 2018. A neural layered model for nested named entity recognition. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1446–1459.

Leippold, M., 2023. Sentiment spin: Attacking financial sentiment with GPT-3. Finance Res. Lett. 55, 103957.

Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., Fung, P., 2021. Crossner: Evaluating cross-domain named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, (15), pp. 13452–13460.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., Wu, H., 2022. Unified structure generation for universal information extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5755–5772.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L., 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. pp. 11048–11064.

Sang, E.F., De Meulder, F., 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

Shanahan, M., McDonell, K., Reynolds, L., 2023. Role play with large language models. Nature 623 (7987), 493–498.

Shen, Y., Ma, X., Tan, Z., Zhang, S., Wang, W., Lu, W., 2021. Locate and label: A two-stage identifier for nested named entity recognition. arXiv preprint arXiv:2105.06804.

Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. Adv. Neural Inf. Process. Syst. 30.

Sohrab, M.G., Miwa, M., 2018. Deep exhaustive model for nested named entity recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2843–2849.

Tan, C., Qiu, W., Chen, M., Wang, R., Huang, F., 2020. Boundary enhanced neural span classification for nested named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, (05), pp. 9016–9023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Vemprala, S., Bonatti, R., Bucker, A., Kapoor, A., 2023. Chatgpt for robotics: Design principles and model abilities. arXiv preprint arXiv:2306.17582.

Wang, J., Shou, L., Chen, K., Chen, G., 2020. Pyramid: A layered model for nested named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5918–5928.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824–24837.

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., Qiu, X., 2021. A unified generative framework for various NER subtasks. 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). pp. 5808–5822.

Yang, Y., Katiyar, A., 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, pp. 6365–6375.

Yang, L., Yuan, L., Cui, L., Gao, W., Zhang, Y., 2022. FactMix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 5360–5371.

Zhang, B., Haddow, B., Birch, A., 2023a. Prompting large language model for machine translation: a case study. In: Proceedings of the 40th International Conference on Machine Learning. pp. 41092–41110.

Zhang, S., Shen, Y., Tan, Z., Wu, Y., Lu, W., 2022. De-bias for generative extraction in unified NER task. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 808–818.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A., 2023b. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923.

Zheng, C., Cai, Y., Xu, J., Leung, H., Xu, G., 2019. A boundary-aware neural model for nested named entity recognition. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics.