JUSTC

Article

http://justc.ustc.edu.cn

Received: August 31, 2023; Accepted: December 04, 2023

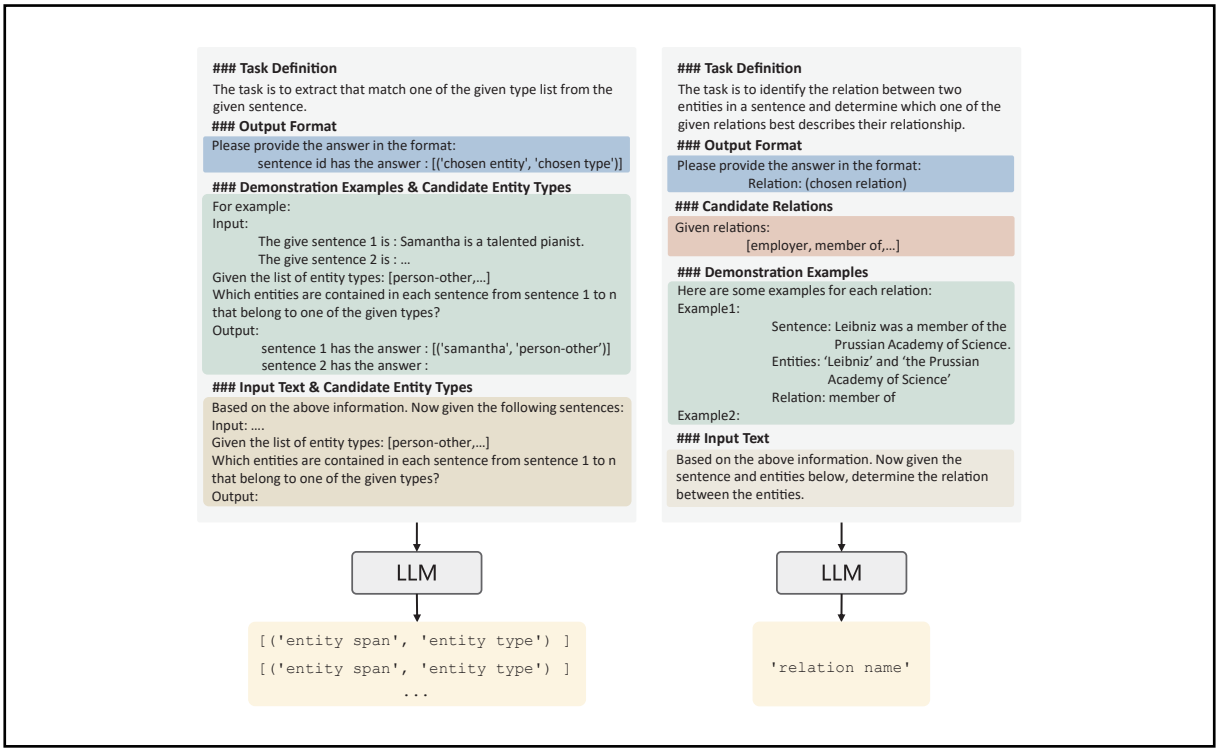
Exploration of augmented prompting methods for information extraction using large language models

Yishuo Fu[1], Benfeng Xu[2] ✉, Mingxuan Du[1], Quan Wang[3], and Zhendong Mao[2]

[1]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China;*
[3]*MOE Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing 100876, China*

✉Correspondence: Benfeng Xu, E-mail: benfeng@mail.ustc.edu.cn

Graphical abstract



The overall framework of the augmented prompting method for information extraction using large language models.

Public summary

- Validating large language models' (LLMs') potential in information extraction: This study systematically evaluates the capability of LLMs in performing information extraction tasks via prompt strategies, providing the first empirical evidence for their practical application in this field.

- Pioneering structured interactive prompting methods: We propose three novel prompting methods (SFP/SIRP/VESIRP) tailored for information extraction, leveraging structured guidance and interactive reasoning to enhance prediction accuracy. VESIRP further introduces a voting mechanism for robust decision-making.

- Achieving breakthroughs in few-shot & zero-shot learning: On the FewNERD benchmark, our approach achieves competitive results on the FewNERD benchmark for both zero-shot and few-shot named entity recognition, offering an efficient solution for low-resource information extraction.

Citation: Fu Y S, Xu B F, Du M X, et al. Exploration of augmented prompting methods for information extraction using large language models. *JUSTC*, **2025**, 55(7): 0702. CSTR: 32290.14.JUSTC-2023-0128. DOI: 10.52396/JUSTC-2023-0128

JUSTC

Article

http://justc.ustc.edu.cn

Received: August 31, 2023; Accepted: December 04, 2023

Exploration of augmented prompting methods for information extraction using large language models

Yishuo Fu[1], Benfeng Xu[2] ✉, Mingxuan Du[1], Quan Wang[3], and Zhendong Mao[2]

[1]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230026, China;*
[3]*MOE Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing 100876, China*

✉Correspondence: Benfeng Xu, E-mail: benfeng@mail.ustc.edu.cn

Cite This: *JUSTC*, **2025**, 55(7): 0702 (10pp)

Read Online

Abstract: Information extraction (IE) aims to automatically identify and extract information about specific interests from raw texts. Despite the abundance of solutions based on fine-tuning pretrained language models, IE in the context of few-shot and zero-shot scenarios remains highly challenging due to the scarcity of training data. Large language models (LLMs), on the other hand, can generalize well to unseen tasks with few-shot demonstrations or even zero-shot instructions and have demonstrated impressive ability for a wide range of natural language understanding or generation tasks. Nevertheless, it is unclear, whether such effectiveness can be replicated in the task of IE, where the target tasks involve specialized schema and quite abstractive entity or relation concepts. In this paper, we first examine the validity of LLMs in executing IE tasks with an established prompting strategy and further propose multiple types of augmented prompting methods, including the structured fundamental prompt (SFP), the structured interactive reasoning prompt (SIRP), and the voting-enabled structured interactive reasoning prompt (VESIRP). The experimental results demonstrate that while directly promotes inferior performance, the proposed augmented prompt methods significantly improve the extraction accuracy, achieving comparable or even better performance (e.g., zero-shot FewNERD, FewNERD-INTRA) than state-of-the-art methods that require large-scale training samples. This study represents a systematic exploration of employing instruction-following LLM for the task of IE. It not only establishes a performance benchmark for this novel paradigm but, more importantly, validates a practical technical pathway through the proposed prompt enhancement method, offering a viable solution for efficient IE in low-resource settings.

Keywords: prompt learning; natural language processing; few-shot information extraction; zero-shot information extraction

CLC number: TP391 **Document code:** A

1 Introduction

In recent years, the field of natural language processing (NLP) has undergone significant research advancements, with information extraction (IE) garnering substantial attention as one of the key NLP tasks. IE aims to extract structured information, such as named entities[1] and entity relationships[2], from unstructured texts. However, traditional IE methods typically require a large amount of annotated data for model training, and in real-world applications, collecting and annotating such large-scale datasets is a time-consuming and costly task. Additionally, real-world scenarios often involve few or even zero-shot situations, where information extraction needs to be performed with very limited samples. Consequently, the introduction of few-shot and zero-shot information extraction methods has become highly necessary.

Few-shot IE aims to identify novel named entities[3] or entity relations[4] on the basis of limited sample data, i.e., support examples. On the other hand, zero-shot IE aims to recognize novel named entities[5] or entity relations[6] without any labeled examples. Since zero-shot IE involves minimal human intervention, it poses even greater challenges than does few-shot information extraction. Investigating these issues can help overcome limitations such as data scarcity and high annotation costs, further advancing the widespread application of natural language processing techniques in practical scenarios.

In recent years, transfer learning[7], meta-learning[8], and zero-shot learning[9,10] methods have been widely applied to address the challenges of few-shot and zero-shot IE tasks. These methods involve pretraining on large-scale data and then fine-tuning the pretrained language model to adapt its knowledge to few-shot or zero-shot tasks, aiming to enhance information extraction performance. However, due to the scarcity of training data, information extraction in few-shot and zero-shot scenarios remains challenging. Additionally, the large language models (LLMs) have demonstrated impressive generalizability in various natural language understanding[11] and generation[12] tasks, even when unseen tasks with few-shot demonstrations or zero-shot instructions are handled. However, the applicability of such effectiveness to IE tasks remains uncertain. IE tasks differ from traditional

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

natural language understanding tasks in which they require accurate recognition and extraction of information in complex and diverse contexts. Moreover, in few-shot and zero-shot IE scenarios, comprehensive assessments are required for model learning, generalization, reasoning, and inductive abilities. In conclusion, while LLMs excel in broad natural language understanding and generation tasks, replicating their effectiveness in IE tasks remains a challenge.

Researches[13–15] have shown that combining LLMs with prompt methods performs well in various downstream NLP tasks. However, a significant performance gap still exists compared with the state-of-the-art (SOTA) results. In this paper, we first validate the effectiveness of LLMs in executing few-shot and zero-shot IE tasks via established prompting strategies[16]. However, owing to the complexity of such tasks, using LLMs directly results in mismatched and inaccurate outputs, which fail to achieve desirable results. To address this issue, we propose and optimize three methods. First, we employ the structured fundamental prompt (SFP) to provide task-specific structured information, guiding LLMs to better understand context and semantics, improving performance. Second, we utilized the structured interactive reasoning prompt (SIRP) to increase the model's reasoning ability through multi-turn conversations, further enhancing its performance. Finally, we introduced the voting-enabled structured interactive reasoning prompt (VESIRP) to increase model robustness and structural consistency and reliability through a multi-round voting mechanism. These methods aim to enhance LLMs performance in few-shot and zero-shot IE tasks. The experimental results demonstrate that these prompt methods achieve comparable or even superior performance to SOTA methods requiring large-scale training samples (e.g., zero-shot FewNERD[17], FewNERD-INTRA). Through in-depth research and comparison of these prompt methods, we aim to bridge the performance gap between LLM and SOTA in few-shot and zero-shot IE tasks, providing robust theoretical and empirical evidence for improving the performance of few-shot and zero-shot IE tasks. The main contributions of our work are summarized as follows:

(Ⅰ) This study validates the effectiveness of large language models (LLMs) in executing IE tasks via established prompt strategies, demonstrating their potential application in the field of information extraction.

(Ⅱ) To enhance the performance of the model in IE tasks, we propose various prompt methods, including structured fundamental prompt (SFP), structured interactive reasoning prompt (SIRP), and voting-enabled structured interactive reasoning prompt (VESIRP). These prompt methods are designed to accommodate the characteristics of information extraction tasks and provide guidance for accurate prediction and extraction.

(Ⅲ) The experimental results demonstrate that our methods achieve SOTA performance in zero-shot named entity recognition on the FewNERD dataset and in few-shot named entity recognition on the FewNERD INTRA dataset. This indicates the significant advantages of our methods in addressing few-shot and zero-shot IE tasks, yielding highly competitive results.

2 Related works

In this section, we briefly summarize the related work in various aspects.

Few-shot information extraction. Few-shot IE refers to the concept of performing information extraction tasks with a limited number of samples (usually only a few or even just one example). In the early stages of research, transfer learning[7] and meta-learning[8] provided crucial avenues for the development of few-shot IE. These methods involve pretraining models on large-scale datasets and then fine-tuning them to adapt to tasks with scarce samples, resulting in initial successes. Pretrained language models (LMs)[18], such as BERT[19], GPT[20,21], and RoBERTa[22], are pretrained with self-supervised[23] objectives and subsequently fine-tuned[24,25] to effectively transfer their learned rich semantic representations to information extraction datasets. Similarly, sequence-to-sequence models, such as T5[26], BART[27], and GLM[28], address few-shot information extraction tasks through techniques like data augmentation and fine-tuning[29].

In the domain of few-shot named entity recognition (NER) for few-shot IE, several approaches have been proposed. ProtoBERT[30] adopts an averaging approach over all samples in the support set for each class to calculate entity prototypes. Classic classification is performed through nearest-neighbor classification on the basis of the the distance between query instances and prototypes. NNShot[31] determines the label of a query instance through the character-level distance. Building upon NNShot, StructShot[31] further enhances classification performance by incorporating an additional Viterbi decoder. CONTaiNER[32] introduces a contrastive learning method aimed at optimizing the distributional distance between labels in few-shot NER tasks. MeTNet[33] introduces the concept of adaptive boundaries and designs an improved triplet network for each entity type to predict labels for query instances. ESD[34] employs various proto-based attention mechanisms to enhance model performance. DecomMETA[35] addresses few-shot span detection and entity type classification sequentially through meta-learning, resolving challenges in few-shot NER. SpanProto[36] transforms sequential labels into a global boundary matrix and utilizes prototype learning to capture semantic representations.

In the domain of few-shot relation extraction (RE) for few-shot IE, several approaches exist. ProtoNet[4] learns a metric space by calculating the distance to prototype representations for each category, facilitating effective classification. The MTB[37] model employs the insertion of markers and positional embeddings around target entities in sentences to label entities involved in the relation to be determined. Furthermore, it replaces entities with a [BLANK] symbol with a fixed probability, directing the model's attention to the analysis of the relationship between the two entities in the sentence. The BERT-PAIR[25] model concatenates two text segments to form an input sequence, which is then encoded by BERT and classified via an additional fully connected layer. MLMAN[38] is a multi-level matching and aggregation network designed for few-shot relation classification tasks, encoding query instances and each support set in an interactive manner.

Zero-shot information extraction. Zero-shot IE refers to

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

performing IE tasks without the presence of sample instances, leveraging the existing knowledge of the model. Initially, methods for zero-shot IE predominantly employed distant supervision[39], and rule-based[40,41] approaches to achieve cross-domain information extraction. However, in recent years, with the rise of pretrained language models such as BERT and GPT, a range of innovative techniques have emerged, including template-based question generation[10,42], generative models[43], and graph neural networks[44]. Moreover, transfer learning strategies such as multi-task learning[45] and meta-learning[46] have been introduced to enhance the model's adaptability in unfamiliar domains by sharing knowledge across multiple tasks. Concurrently, some studies have focused on integrating external information sources[47] such as knowledge graphs to improve zero-shot IE performance. This technology not only holds potential applications in traditional entity-relation extraction but also contributes to information retrieval and question-answering domains.

Prompt learning. Prompt learning aims to guide the learning and reasoning process of models through the design of effective prompts. Its objective is to reduce reliance on a large amount of annotated data by designing appropriate prompts to facilitate learning for specific tasks. The earliest prompt learning methods can be traced back to rule-based system approaches[48]. With the development of deep learning models, there have been emerging techniques such as PET[49], which uses manually designed hard prompts, and methods that leverage automatically generated soft prompts[50]. Following the release of GPT-3.5, various NLP tasks have been tackled using GPT-3.5 combined with prompt learning methods, including zero-shot information extraction by Gao et al.[51] and stance detection by Zhang et al[52].

3 Task definition

In this study, we focus on the tasks of NER and RE in the field of IE. NER aims to accurately identify and extract the named entities and their corresponding predefined entity types in the sentence. On the other hand, RE aims to accurately identify and extract predefined relation types between given entity pairs.

Specifically, we adopt the N-way K-shot setting for both few-shot NER and few-shot RE in this study. An example of a 2-way 1-shot NER and RE problem is shown in Fig. 1. In each input task, there is a support set and a query set. The support set serves as the learning data for the model, providing N predefined types with K-labeled samples for each type. The query set contains samples that the model needs to recognize and predict.

4 Methods

This paper proposes several enhanced prompt methods, including the structured fundamental prompt (SFP), the structured interactive reasoning prompt (SIRP), and the voting-enabled structured interactive reasoning prompt (VESIRP), in addition to established prompt strategies to improve the performance of LLMs in few-shot and zero-shot IE tasks.

4.1 Structured fundamental prompt

On the basis of preliminary experimental results, we find that LLMs enhances the performance of IE tasks by learning rich semantic knowledge. Considering the requirements of input and output for IE tasks and the characteristics of LLMs, we devised a structured fundamental prompt (SFP). This prompt integrates various components, such as task definition, candidate categories, demonstration samples, output formats, and input texts, via natural language expressions. Through this integration approach, we aim to better meet the task requirements and leverage the capabilities of LLMs to improve the effectiveness of information extraction. Fig. 2 presents a straightforward example of the SFP employed for both few-shot NER and few-shot RE.

For the SFP of few-shot NER, the task is defined as

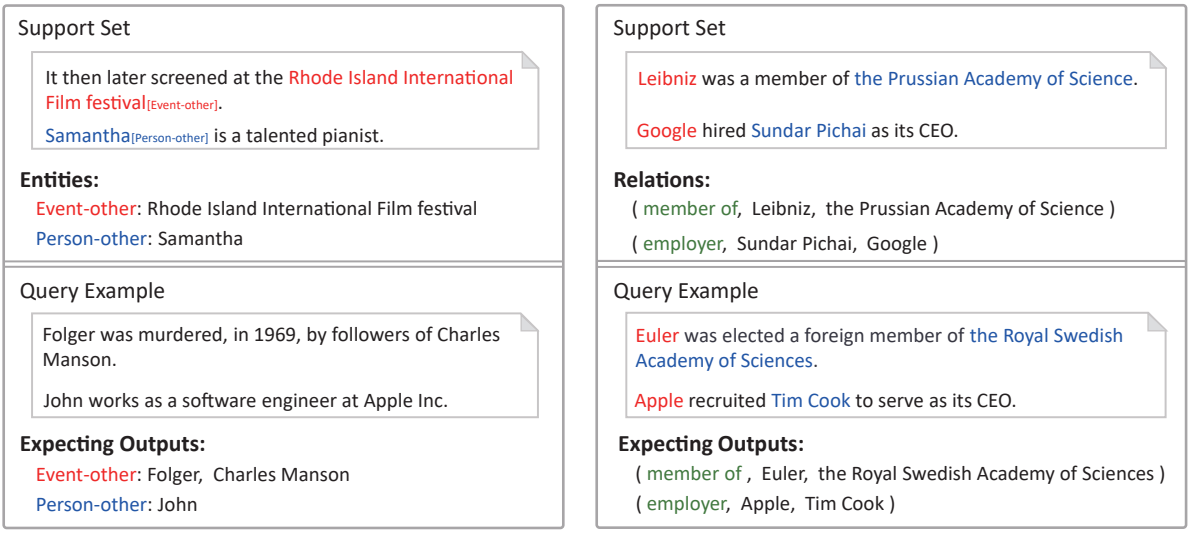


Fig. 1. Examples of a 2-way 1-shot IE task. On the left, it demonstrates a 2-way 1-shot NER task, where a support set with two target entity types, each linked to a single labeled entity, is provided. The aim is to identify entities within a query set. On the right, it showcases a 2-way 1-shot RE task. Different colors represent distinct entities; Red signifies the head entity, and blue denotes the tail entity. A support set contains two target relations, each associated with a single entity pair. The goal is to identify the relationships between specified entity pairs in the query set.

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

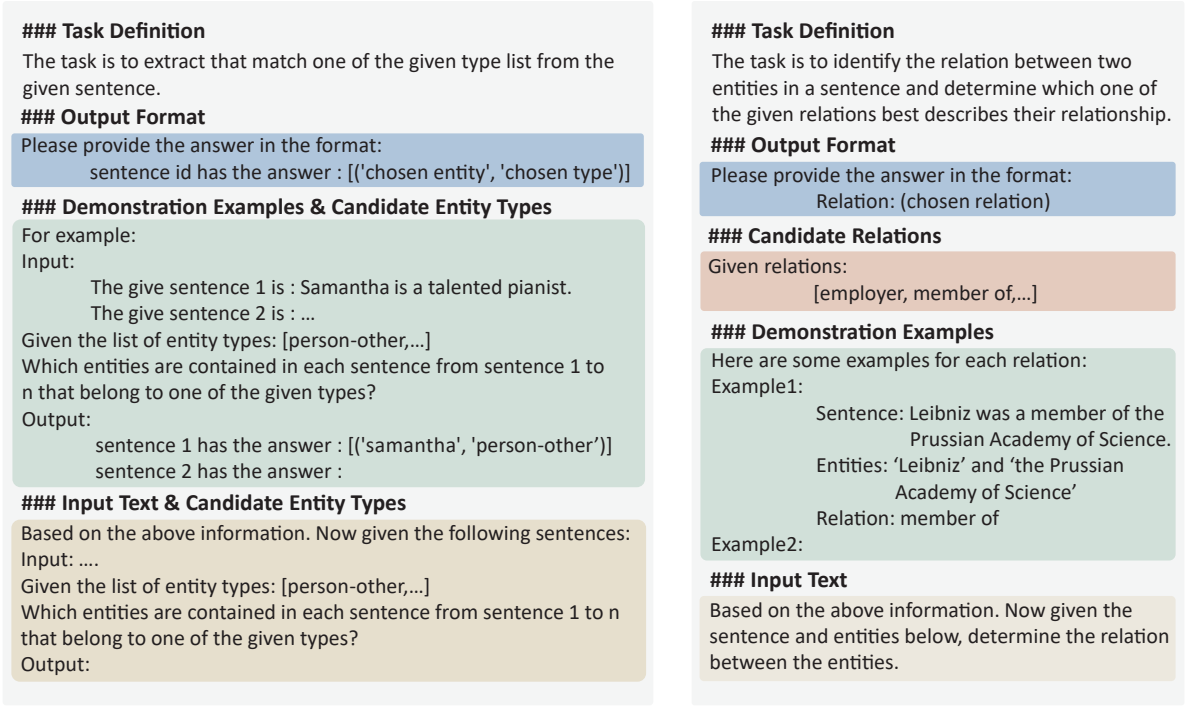


Fig. 2. Examples of a structured fundamental prompt (SFP) for few-shot IE, the left side presents an SFP example for few-shot NER, while the right side presents an SFP example for few-shot RE. The examples of SIRP, VESIRP, and SFP share similarities, with specific distinctions elucidated in the details of Sections 4.2 and 4.3 for reference.

identifying entities in sentences that belong to the expected entity type set. The candidate categories represent the expected entity type set, and the demonstration samples consist of the support set information in the few-shot NER. The input text refers to the query set samples. To achieve the goal of NER, the model needs to output the spans of entities and their corresponding entity types. We adopt a question-answering format to construct this SFP, which addresses the complexity of NER. In the design process, we incorporate the candidate entity types into the premise of the question-answering format to help the LLM better understand the task objective. Additionally, since each task in our dataset has a large number of support set and query set samples, and after multiple experiments, we have observed that the LLM has difficulty in outputting results in the input order. To address this issue, we introduce natural language numbering, which requires the LLM to output results according to the assigned numbers, ensuring that each input sentence has a corresponding output. The left portion of Fig. 2 illustrates a simple example of this task.

For the SFP of zero-shot NER, we modify the SFP of few-shot NER by removing the demonstration samples, aligning with the requirements of zero-shot NER.

Furthermore, many methods for few-shot NER that are based on fine-tuning pretrained language models address the label dependency issue by decomposing NER into two subtasks: entity span extraction and entity classification. In this study, we delve deeper into the effectiveness of this decomposition strategy and its implementation through the use of prompts. Notably, most fine-tuning methods for pretrained language models approach few-shot NER in the order of "extract entity spans first, then perform entity classification". However, our proposed methods, which leverage LLMs combined with prompt learning, challenge this traditional order by first identifying potential entity types in a given sentence and subsequently extracting the corresponding entity spans. Hence, the SFP for few/zero-shot NER encompasses the following two novel approaches:

(Ⅰ) Entity span extraction followed by candidate entity classification (SFP-1).

(Ⅱ) Identification of entity types followed by extraction of the corresponding entity spans (SFP-2).

These two prompts merely involve changing the task objective and the questioning approach on the basis of the SFP of few/zero-shot NER. Fig. 3 provides a simple example of this method.

For the SFP of few-shot RE, the task is defined as determining the specific relation type among a given set of relation types for a given entity pair in a sentence. The candidate categories represent the expected set of relationship types, and concrete relationship names are used instead of abstract relationship labels in the dataset. The support set in the few-shot RE serves as a demonstration sample to provide the LLM with knowledge of the relationship types. To standardize the output format of the model and facilitate the learning of relationship types, we employ a format similar to "Sentence: ... Entities: ... Relation: ..." in the demonstration samples. The query set samples in the few-shot RE are used as input texts. The right portion of Fig. 2 shows a simple example of this task.

For the SFP of zero-shot RE, the premise is the absence of any learnable samples, so the demonstration samples in the SFP of few-shot RE can be removed directly to serve as the

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

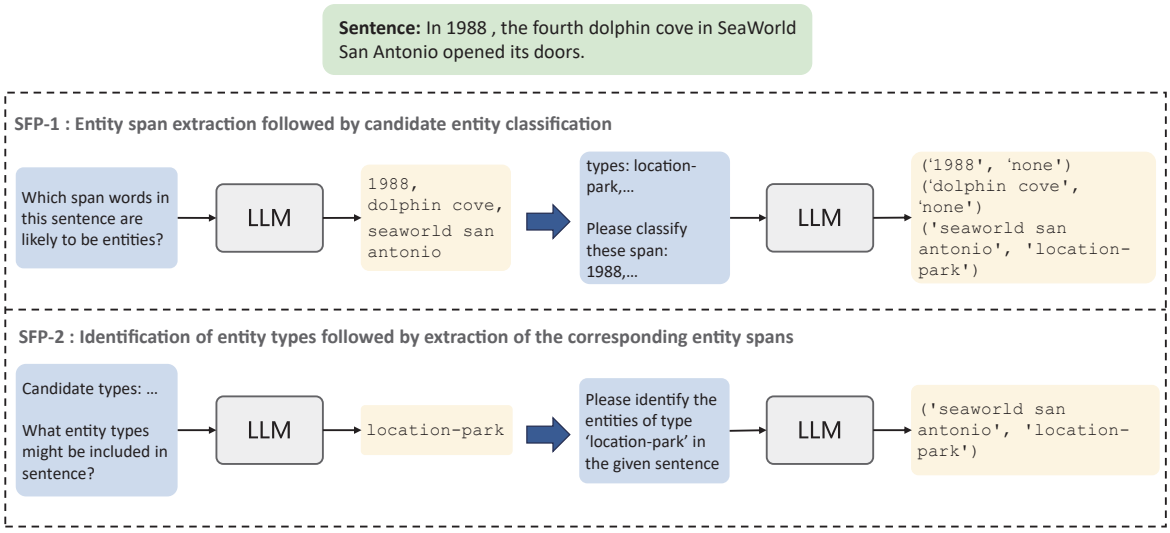


Fig. 3. An example of a structured fundamental prompt (SFP) framework for named entity recognition (NER) employing a two-stage approach is presented. In SFP-1, the initial stage involves extracting the span of the entity followed by classification. Conversely, in SFP-2, the process begins with identifying the category and subsequently extracting the corresponding span.

structured fundamental prompt for zero-shot RE.

4.2 Structured interactive reasoning prompt

With the increase in model size, researches have shown that the inference capability of the model also improves. To further enhance the performance of few/zero-shot NER and few/zero-shot RE, we propose the structured interactive reasoning prompt (SIRP).

Inspired by the concept of the chain of thought[53], our SIRP introduces an inference process on top of the existing SFP, requiring the model to generate a series of reasoning steps before the final output (such as relation, entity, span, and type) is selected. To achieve this goal, we incorporate the corresponding inference process into the learnable samples based on the SFP framework depicted in Fig. 2. However, manually constructing coherent reasoning processes is complex and time-consuming. To address this issue, we propose an iterative dialog approach that allows the model to generate the reasoning process automatically. The generated reasoning process is then included in the learnable samples, enhancing the model's ability to learn the reasoning process. Taking RE as an example, Fig. 4 illustrates a simple example of generating the reasoning process through iterative dialog.

4.3 Voting-enabled structured interactive reasoning prompt

In generative large-scale language models, the temperature parameter is an important factor that controls the randomness of text generation. Higher temperature values make the generated text more random, resulting in less stable outputs, but also providing a wider range of experimental results. Considering the potential inaccuracy in single-step inference, we propose a voting-enabled structured interactive reasoning prompt (VESIRP) based on a voting mechanism to increase the accuracy and stability of the generated output in few/zero-shot NER and few/zero-shot RE.

VESIRP is a natural language-based prompting method designed for information extraction tasks. It integrates various components to execute the information extraction task, including task definitions, candidate categories, learnable samples, inference processes, output formats, and input text. A key feature of VESIRP is its utilization of temperature parameters to modulate the model's generation process. This involves the model generating multiple possible answers by increasing the model temperature parameter and subsequently selecting the answer with the highest frequency from these generated results as the final output. Specifically, VESIRP employs SIRP initially to guide the model in generating text, allowing the model to generate multiple possible answers by increasing the model temperature parameter. The frequency of these answers is then statistically evaluated, and the answer with the highest occurrence is chosen as the final output. Through this design, we can maintain a certain level of randomness while improving the accuracy of the generated output, thereby enhancing the stability and reliability of the model.

5 Experiments

5.1 Datasets

For few/zero-shot NER, we employ the FewNERD[17] dataset. The FewNERD dataset is specifically designed for the task of named entity recognition and encompasses 8 coarse-grained and 66 fine-grained entity types. Each entity label includes hierarchical information for both coarse-grained and fine-grained structures. The dataset comprises over 180000 sentences, more than 490000 annotated entities, and over 4.6 million characters. FewNERD is divided into two parts: FewNERD(INTRA) and FewNERD(INTER). In the FewNERD(INTRA) dataset, different sets contain entities of different coarse-grained types. In the FewNERD(INTER) dataset, three sets include entities of the same coarse-grained type but different fine-grained types.

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

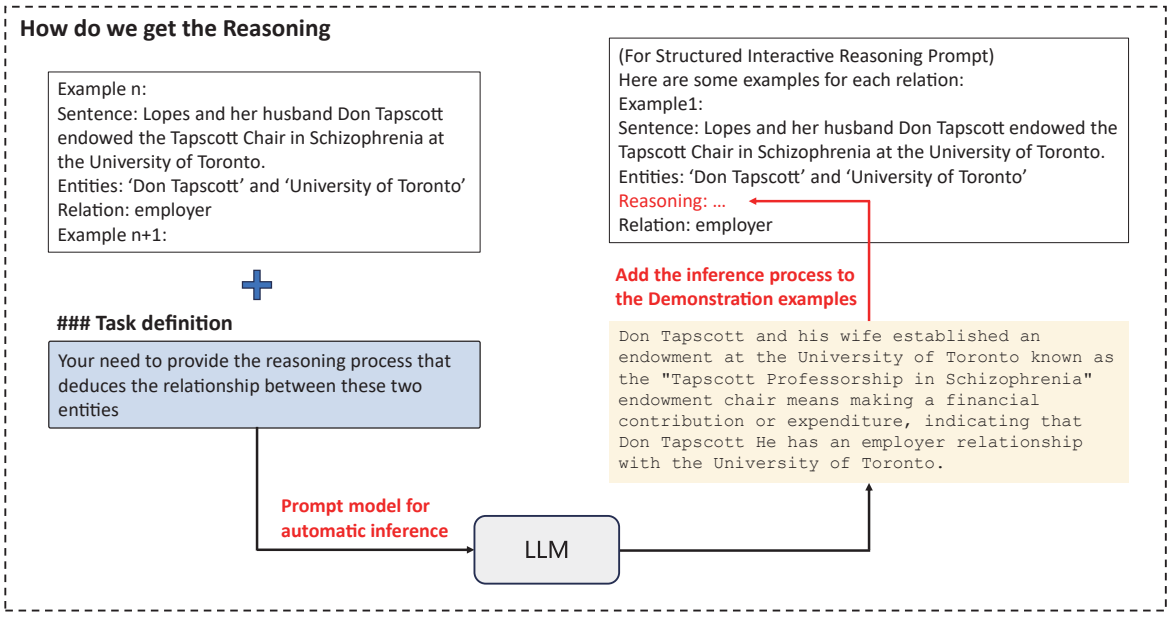


Fig. 4. An example of reasoning generation for RE.

For few/zero-shot RE, we employ the FewRel[4] dataset. FewRel is a benchmark dataset designed for relation classification tasks, with a specific emphasis on few-shot learning scenarios. The dataset is designed to evaluate whether models can accurately classify relationships in text when provided with extremely limited examples. It covers a diverse set of relationship categories, including person relations, organizational relations, among others, totaling 100 distinct relationship categories, each with 700 instances.

5.2 Evaluation metrics

We consider only exact matches and use micro-F1 to evaluate the NER task. We consider a prediction to be correct only when both the entity span and the entity type match the ground truth entity.

We adopt accuracy as the evaluation metric for relation extraction. Accuracy represents the proportion of correctly predicted relation samples to the total number of samples.

6 Results and discussion

In this study, we adopted the established prompting strategy from Ref. [16] as the primary baseline prompt (BP). Considering the input token limitations of the GPT-3.5 model, we only conducted necessary comparative experiments. We evaluated the effectiveness of the three prompting methods, SFP, SIRP, and VESIRP, separately in few-shot NER and few-shot RE tasks. Based on the performance of these prompting methods in few-shot tasks, we determined which method to validate in zero-shot NER and zero-shot RE tasks.

6.1 Results of few/zero-shot NER

Table 1 presents the results of few-shot NER. For comparison, we selected several representative models on the FewNERD dataset, including ProtoBERT, NNShot, StructShot, CONTaiNER, MeTNet, ESD, DecomMeta and SpanProto. These methods are all based on fine-tuning pretrained language models. ProtoBERT, NNShot, StructShot, CONTaiNER, and MeTNet obtain entity spans and their corresponding entity categories in one stage. On the other hand, ESD, DecomMeta, and SpanProt divide few-shot NER into two stages: extracting entity spans first and then classifying these entity spans.

Based on the preliminary experimental results of few-shot NER as shown in Table 1, we specifically validated BP, SFP, SFP-1, SFP-2, and the inference-stimulating VESIRP, under the zero-shot scenario on the FewNERD dataset. The results of zero-shot NER are presented in Table 2.

6.2 Discussion of few/zero-shot NER

As shown in Table 1, SFP performs the best in few-shot NER, whereas SFP-1 shows the weakest performance, although it is still significantly better than BP. In particular, SFP demonstrates superior performance on FewNERD(INTRA) and outperforms existing SOTA models significantly. The performance of SFP on FewNERD(INTER) is slightly better than that of one-stage models but inferior to that of existing two-stage models, with a notable gap. Notably, the performance of SFP on FewNERD(INTRA) and FewNERD(INTER) is similar, with the main difference being that existing methods based on fine-tuning pretrained models perform poorly on FewNERD(INTRA). This is because all entities in the training, validation, and test sets of FewNERD(INTRA) belong to different coarse-grained types, making it challenging for fine-tuning-based pretraining models to generalize well. In contrast, LLM, with its rich knowledge and strong generalization ability, exhibits comparable performance on both the FewNERD(INTRA) and FewNERD(INTER) datasets.

Among SFP, SFP-1, and SFP-2, the one-stage SFP performs the best, where as the two-stage SFP-1 performs the worst, with SFP-2 coming in second after SFP. Although

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

Table 1. F1 scores (%) of 5-way 1-shot (5-1), 5-way 5-shot (5-5), 10-way 1-shot (10-1), and 10-way 5-shot (10-5) problems on few-shot NER over FEWNERD dataset.

Methods	FewNERD (INTRA)				FewNERD (INTER)			
	5-1	5-5	10-1	10-5	5-1	5-5	10-1	10-5
ProtoBERT	23.45±0.92	41.93±0.55	19.76±0.59	34.61±0.59	44.44±0.11	58.80±1.42	39.09±0.87	53.97±0.38
NNShot	31.01±1.21	35.74±2.36	21.88±0.23	27.67±1.06	54.29±0.40	50.56±3.33	46.98±1.96	50.00±0.36
StructShot	35.9±0.69	38.8±1.72	25.3±0.84	26.3±2.59	57.3±0.53	57.1±2.09	49.46±0.53	49.39±1.77
CONTaiNER	40.43	53.70	33.84	47.49	55.95	61.83	48.35	57.12
MeTNet	55.79±0.23	65.41±0.35	47.18±0.89	60.71±0.17	74.42±0.61	76.28±0.32	67.91±0.68	71.96±0.35
ESD	41.44±1.16	50.68±0.94	32.29±1.10	42.92±0.75	66.46±0.49	74.14±0.80	59.95±0.69	67.91±1.41
DecomMeta	52.04±0.44	63.23±0.45	43.50±0.59	56.84±0.14	68.77±0.24	71.62±0.16	63.26±0.40	68.32±0.10
SpanProto	54.49±0.39	65.89±0.82	45.39±0.72	59.37±0.47	73.36±0.18	75.19±0.77	66.26±0.33	70.39±0.63
BP	36.49	32.98	34.29	25.54	29.98	25.27	26.27	21.69
SFP (Us)	**64.95**	**67.55**	**60.47**	**61.23**	61.88	**67.06**	**61.87**	**64.26**
SFP-1 (Us)	44.53	53.76	–	–	42.00	52.47	–	–
SFP-2 (Us)	62.05	63.82	53.27	–	**63.37**	64.83	57.98	–
SIRP (Us)	52.78	59.58	–	–	49.21	57.49	–	–
VESIRP (Us)	60.32	–	–	–	58.75	–	–	–

The best performance of the proposed method is highlighted in bold.

Table 2. F1 scores (%) of 5-way, 10-way problems on zero-shot NER over FewNERD dataset.

Methods	FewNERD (INTRA)		FewNERD (INTER)	
	5-way	10-way	5-way	10-way
BP	25.69	18.29	21.26	13.92
SFP (Us)	**58.48**	**56.12**	**58.33**	**55.17**
SFP-1 (Us)	35.04	32.73	30.53	25.06
SFP-2 (Us)	55.58	44.29	56.80	45.69
VESIRP (Us)	56.72	50.07	49.56	44.05

The best performance of the proposed method is highlighted in bold.

entity span extraction and entity classification are relatively simpler than NER for LLMs, the issue of error propagation between two-stage tasks cannot be ignored. In particular, the problem of error propagation is more prominent in SFP-1. This is attributed to the fact that, in the first stage, SFP-1 extracts entity spans from the text with the sole requirement that the output spans are entities. However, LLMs tend to extract numerous candidate entity spans, many of which do not belong to the predefined set of entity categories. These candidate entity spans outside the predefined categories pose challenges in the second stage of entity classification, as they are difficult to filter effectively. Conversely, LLMs are more inclined to categorize these candidate entity spans as one of the predefined categories. Consequently, SFP-1 exhibits suboptimal performance in named entity recognition. In contrast, SFP-2 constrains the scope of entity type recognition in the first stage, requiring the LLM to identify only those entity types within the predefined set of entity categories present in the text. This approach significantly improves the accuracy of LLM outputs, enhancing precision during the second stage of entity span extraction for the corresponding entity types. While potential issues of error propagation may persist, the performance of SFP-2 has shown a notable improvement over that of SFP-1.

Given the outstanding performance of the SFP method, both the SIRP and VESIRP approaches are enhancements built upon the SFP framework. The SIRP and VESIRP methods aim to stimulate the model's reasoning capabilities to improve predictive performance. In theory, they should outperform SFP in terms of performance. However, as indicated by the data in Table 1, the performance of SFP still surpasses that of SIRP and VESIRP, which contradicts the expected performance improvement. This discrepancy arises from the fact that in the N-way K-shot tasks of the FewNERD dataset, each entity type includes at least K samples in the query set for model identification. In few-shot NER, the initial number of input tokens is already substantial, and the inference processes for learnable examples provided in the support set further significantly increase the number of input tokens. Consequently, this may lead to errors or information omission by LLM when comprehending query samples. The experimental results also confirm that when facing complex queries that LLM finds challenging to comprehend, it is more likely to produce empty outputs. Compared with the SIRP, the VESIRP improves performance significantly by employing a five-round voting mechanism to filter out empty outputs. Nevertheless, the fundamental issue of input length remains, causing the VESIRP to still fall short in performance compared with SFP.

As shown in Table 2, we observe that SFP performs the best in the zero-shot scenario, and its performance also surpasses that of VESIRP. This can be attributed to the fact that in the zero-shot context, no learnable examples or corresponding inference processes are provided to the LLM. The absence of learnable examples to stimulate the reasoning capabilities of LLM results in LLM being unable to acquire new information from examples. Instead, it relies solely on its

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

internal reasoning processes to complete the task. This situation increases the complexity of the task, as LLM needs to independently reason and produce answers without relying on known examples for cues. This further demonstrates that the reasoning ability of LLMs needs to be effectively stimulated through the provision of reasoning samples. However, it is worth noting that the zero-shot performance of SFP on the FewNERD dataset is also superior to that of several existing few-shot methods.

6.3 Results of few/zero-shot RE

Table 3 presents the results of few-shot RE. For comparison, we selected representative existing approaches on the FewRel dataset, including MTB, BERT-PAIR, MLMAN, and ProtoNet. These approaches are based on fine-tuning pretrained language models. Among them, the MTB achieved the best performance on the FewRel dataset. Additionally, we also compared the performance of the human annotators on the dataset.

Based on the preliminary experimental results of few-shot RE as shown in Table 3 and the unique representation of relation labels in the FewRel dataset (using specific symbolic numbers instead of label names), we specifically validated SFP and the optimal method VESIRP for zero-shot scenarios in the FewRel dataset. The results of zero-shot RE are presented in Table 4.

6.4 Discussion of few/zero-shot RE

According to the results in Table 3, the VESIRP performs the best in few-shot RE, while SFP performs relatively poorly. Nevertheless, SFP still outperforms BP by a significant margin and shows comparable performance to other methods, except for the SOTA method. The performance of SFP, SIRP, and VESIRP on FewRel aligns with expectations because of the N-way K-shot nature of tasks in the FewRel dataset. In these tasks, the query set comprises only one relationship sample used for model prediction. Introducing the inference process for learnable examples does not substantially increase the input token length and does not impact the understanding capability of LLMs regarding query samples. In contrary, it effectively stimulates the reasoning capabilities of LLMs. Comparing the effects of the SIRP and SFP, it is evident that the inclusion of an inference process enhances the model's reasoning ability and thereby improves task performance, albeit with limited gains. For the VESIRP, we employ a five-round voting mechanism, which enhances accuracy on the FewRel dataset compared with SIRP and SFP, reducing misclassifications in single outputs. Comparing the results between the 1-shot and 5-shot scenarios, we observe that for LLMs, a single learnable sample is sufficient to acquire rich knowledge. In contrast, the performance gain from multiple learnable samples is not significant. However, for methods employing fine-tuning pretrained models, a larger number of learnable samples leads to better performance in few-shot scenarios.

As shown in Table 4, SFP performs significantly better than VESIRP in zero-shot RE. This may be attributed to the fact that the reasoning capabilities of LLMs require effective stimulation from inferential samples. In the case of zero-shot RE tasks, where there are no learnable samples to stimulate LLM's reasoning abilities, LLM is unable to learn from the samples and only needs to output its own reasoning process. This situation can introduce complexity to the task. Consequently, in zero-shot RE, the LLM performs less favorably on the SIRP and VESIRP than does the SFP. However, it is worth noting that the zero-shot prompt effectiveness of SFP on the FewRel dataset also outperforms certain existing few-shot methods.

Additionally, existing methods based on fine-tuning pretrained language models require training a new model separately. In contrast, the combination of prompt learning with LLM does not involve fine-tuning or parameter updating, significantly reducing computational and time costs. Leveraging these advantages, the performances of SFP, SIRP, and VESIRP still surpass those of methods that are based on fine-tuning pretrained language models.

Table 3. Accuracy (%) of 5-way 1-shot (5-1), 5-way 5-shot (5-5), 10-way 1-shot (10-1), and 10-way 5-shot (10-5) problems on few-shot RE over FewRel dataset.

Methods	5-1	5-5	10-1	10-5
Human	92.22	–	85.88	–
SOTA (MTB)	93.86	97.06	89.20	94.27
BERT-PAIR	88.32	93.22	80.63	87.02
MLMAN	82.98	92.66	73.59	87.29
ProtoNet	81.40	92.11	72.51	86.03
BP	48.14	79.39	58.11	66.44
SFP (Us)	89.88	90.96	81.90	82.40
SIRP (Us)	90.20	91.50	–	–
VESIRP (Us)	**92.47**	**93.81**	**86.19**	–

The best performance of the proposed method is highlighted in bold.

Table 4. Accuracy (%) of 5-way, 10-way problems on zero-shot RE over FewRel dataset. The best performance of the proposed method is highlighted in bold.

Methods	5-way	10-way
SFP (Us)	**85.48**	**76.25**
VESIRP (Us)	82.29	71.14

7 Conclusions

In this study, we first validate the effectiveness of large-scale pretrained language models in executing information extraction (IE) tasks via established prompting strategies. We further propose multiple types of enhanced prompting methods to improve LLM's performance in few-shot and zero-shot information extraction tasks through prompting learning. Specifically, we propose the structured fundamental prompt (SFP) which provides task-specific structured information to guide the LLM in better understanding the task elements. Furthermore, we introduce the structured interactive reasoning prompt (SIRP) to foster the model's reasoning ability, resulting in further performance improvement. Finally, we propose the voting-enabled structured interactive reasoning prompt (VESIRP), which leverages a multi-round voting mechanism

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

to enhance output consistency and reliability. We validated the effectiveness of these prompt methods on FewNERD and FewRel datasets. The experimental results show that these methods achieve comparable or even superior performance to SOTA methods requiring large-scale training samples (e.g., zero-shot FewNERD, few-shot FewNERD-INTRA). This is one of the first studies that address the task of IE in the paradigm of instruction-following LLM, which establishes benchmark results as well as augmented improvements to elicit more future studies. In the future, we will explore additional prompt methods and extend our approach to other task domains.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62222212).

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yishuo Fu is currently pursuing a master's degree at the School of Information Science and Technology, University of Science and Technology of China. Her research mainly focuses on natural language processing and information extraction.

Benfeng Xu is currently pursuing a Ph.D. degree in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. His research mainly focuses on large language models and their versatile applications.

References

[1] Mohit B. Named entity recognition. In: Zitouni I. editor. Natural Language Processing of Semitic Languages. Theory and Applications of Natural Language Processing. Berlin, Germany: Springer, **2014**: 221–245.

[2] Bach N, Badaskar S. A review of relation extraction. **2007**. https://www.bibsonomy.org/bibtex/2b53ab6f9505b53603a1f6ef2509-22e81/schwemmlein. Accessed October 10, 2023.

[3] Huang J, Li C, Subudhi K, et al. Few-shot named entity recognition: A comprehensive study. arXiv: 2012.14978v1, **2020**.

[4] Han X, Zhu H, Yu P, et al. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, **2018**: 4803-4809.

[5] Obeidat R, Fern X, Shahbazi H, et al. Description-based zero-shot fine-grained entity typing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota, USA: Association for Computational Linguistics, **2019**: 807–814.

[6] Levy O, Seo M, Choi E, et al. Zero-shot relation extraction via reading comprehension. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Stroudsburg, PA, USA: ACL, **2017**: 333–342.

[7] Torrey L, Shavlik J. Transfer learning. In: Soria E, Martin J, Magdalena R, et al. editors. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. Madison WI, USA: IGI Global, **2010:** 242–264.

[8] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, **2017**: 70.

[9] Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, **2009**: 951–958.

[10] Palatucci M, Pomerleau D, Hinton G E, et al. Zero-shot learning with semantic output codes. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems. New York: ACM, **2009**: 1410–1418.

[11] Chen X, Ye J, Zu C, et al. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. arXiv: 2303.00293, **2023**.

[12] Maddigan P, Susnjak T. Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. *IEEE Access*, **2023**, *11*: 45181–45193.

[13] Madaan A, Tandon N, Clark P, et al. Memory-assisted prompt editing to improve GPT-3 after deployment. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, **2022**: 2833–2861.

[14] Wei X, Cui X, Cheng N, et al. ChatIE: zero-shot information extraction via chatting with ChatGPT. arXiv: 2302.10205, **2023**.

[15] Agrawal M, Hegselmann S, Lang H, et al. Large language models are few-shot clinical information extractors. Abu Dhabi, United Arab Emirates: ACL, **2022**: 1998–2022.

[16] Han R, Peng T, Yang, et al. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. arXiv: 2305.14450, **2023**.

[17] Ding N, Xu G, Chen Y, et al. 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), online: ACL, **2021:** 3198–3213.

[18] Han X, Zhang Z, Ding N, et al. Pre-trained models: Past, present and future. *AI Open*, **2021**, *2*: 225–250.

[19] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota, USA: ACL, **2019:** 4171–4186.

[20] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training, **2018**. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95-b5003d0a5035. Accessed October 10, 2023

[21] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, **2020**: 1877–1901.

[22] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach, arXiv: 1907.11692, **2019**.

[23] Liu X, Zhang F, Hou Z, et al. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, **2023**, *35*: 857–876.

[24] Joshi M, Chen D, Liu Y, et al. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, **2020**, *8*: 64–77.

[25] Gao T, Han X, Zhu H, et al. FewRel 2.0: Towards more challenging few-shot relation classification. 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). Stroudsburg, PA, USA: ACL, **2019**: 6249–6254.

[26] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **2020**, *21* (1): 5485–5551.

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702

[27] Lewis M, Liu Y, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, **2020**: 7871–7880.
[28] Du Z, Qian Y, Liu X, et al. All NLP tasks are generation tasks: A general pretraining framework, arXiv: 2103.10360, **2021**.
[29] Paolini G, Athiwaratkun B, Krone J, et al. Structured prediction as translation between augmented natural languages. In: International Conference on Learning Representations. Vienna, Austria: ICLR, **2021**.
[30] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, **2017**: 4080– 4090.
[31] Yang Y, Katiyar A. 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). Stroudsburg, PA, USA: ACL, **2020**: 6365-6375.
[32] Das S S S, Katiyar A, Passonneau R J, et al. 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). Dublin, Ireland: ACL, **2022:** 6338–6353.
[33] Han C, Zhu R, Kuang J, et al. Meta-learning triplet network with adaptive margins for few-shot named entity recognition. arXiv: 2302.07739, **2023**.
[34] Wang P, Xu R, Liu T, et al. An enhanced span-based decomposition method for few shot sequence labeling. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: ACL, **2022**: 5012–5024.
[35] Ma T, Jiang H, Wu Q, et al. Decomposed meta-learning for few-shot named entity recognition. In: Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg, PA, USA: ACL, **2022**: 1584–1596.
[36] Wang J, Wang C, Tan C, et al. SpanProto: A two-stage span-based prototypical network for few-shot named entity recognition. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: ACL, **2022:** 3466–3476.
[37] Baldini Soares L, FitzGerald N, Ling J, et al. Matching the blanks: Distributional similarity for relation learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, **2019**: 2895–2905.
[38] Ye Z-X, Ling Z-H. Multi-level matching and aggregation network for few-shot relation classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, **2019:** 2872–2881.
[39] Kurfali M, Wirén M. Zero-shot cross-lingual identification of direct speech using distant supervision. In: Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Online: International Committee on Computational Linguistics, **2023:** 105–111.
[40] Rocktäschel T, Singh S, Riedel S. Injecting logical background knowledge into embeddings for relation extraction. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL, **2015**: 1119–1129.
[41] Demeester T, Rocktäschel T, Riedel S. Lifted rule injection for relation embeddings. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, **2016**: 1389–1399.
[42] Ma K, Ilievski F, Francis J M, et al. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **2021**, *35* (15): 13507–13515.
[43] Qin P, Wang X, Chen W, et al. Generative adversarial zero-shot relational learning for knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, **2020**, *34*: 8673–8680.
[44] Geng Y, Chen J, Ye Z, et al. Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs. *Semantic Web*, **2021**, *12*: 741–765.
[45] Ahuja K, Kumar S, Dandapat S, et al. Multi task learning for zero shot performance prediction of multilingual models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: ACL, **2022:** 5454–5467.
[46] Verma V K, Brahma D, Rai P. A meta-learning for generalized zero-shot learning. arXiv: 1909.04344, **2020**.
[47] Chen Y, Jiang H, Liu L, et al. An empirical study on multiple information sources for zero-shot fine-grained entity typing. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: ACL, **2021**: 2668–2678.
[48] Kay M. The proper place of men and machines in language translation. *Machine Translation*, **1997**, *12*: 3–23.
[49] Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: ACL, **2021**: 255–269.
[50] Gao T, Fisch A, Chen D. Making pre-trained language models better few-shot learners. 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). Online: ACL, **2021**: 3816–3830.
[51] Yuan C, Xie Q, Ananiadou S. Zero-shot temporal relation extraction with ChatGPT. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: ACL, **2023:** 92–102.
[52] Zhang B, Ding D, Jing L. How would stance detection techniques evolve after the launch of ChatGPT? arXiv: 2212.14548, **2022**.
[53] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022). New York: Curran Associates, Inc., **2022**, *35*: 22199–22213.

CSTR: 32290.14.JUSTC-2023-0128
DOI: 10.52396/JUSTC-2023-0128
JUSTC, **2025**, 55(7): 0702