Information Extraction in Low-Resource Scenarios: Survey and Perspective

Abstract—Information Extraction (IE) seeks to derive structured information from unstructured texts, often encountering obstacles in low-resource scenarios due to data scarcity and unseen classes. This paper presents a review of neural approaches to low-resource IE from traditional and LLM-based perspectives, systematically organizing them into a fine-grained taxonomy. Our empirical studies compare LLM-based methods with prior leading models, revealing that: (1) well-tuned LMs perform relatively best; (2) tuning open-resource LLMs and in-context learning with GPT family are generally effective; (3) LLMs struggle to tackle complex tasks with intricate schema. Furthermore, we compare traditional methods and discuss LLM-based approaches, spotlighting promising applications and delineating future research directions. This survey aims to foster understanding of this field, inspire new ideas, and encourage widespread applications in both academia and industry.

I. INTRODUCTION

Information Extraction (IE) [1], [2] offers essential support for a diverse range of domain-specific [3], [4] and knowledge-intensive [5], [6] tasks, thus appealing to AI community. Common IE tasks, including Named Entity Recognition (NER) [7], Relation Extraction (RE) [8], and Event Extraction (EE) [9], aim to derive structured data from unstructured texts. While deep learning has boosted IE performance, it demands massive amounts of labeled data, which is difficult to acquire due to task and domain variability in practice. This issue motivates research in low-resource IE (data-efficient IE), which seeks to enhance learning efficiency under real world contexts with sparse data.

Low-resource IE has been widely investigated due to its potential for making models data-efficient and adaptable to various scenarios. Among numerous approaches proposed in recent years, Large Language Models (LLMs) [10], [11] have demonstrated promising performance in low-resource scenarios. Hence, a timely review on low-resource IE is beneficial in offering insights to both researchers and industry practitioners. Though several surveys on task-specific low-resource IE (e.g., NER [12], [13], RE [14], [15], EE [16]) and low-resource NLP [17], [18] have been published, a comprehensive survey for low-resource IE encompassing both traditional models and modern LLMs is still absent. Therefore, in this paper, we provide a literature review on low-resource IE, hoping to systematically analyse the methodologies, compare different technical solutions and inspire diverse new ideas.

In this paper, we begin by introducing the basics of low-resource learning and IE in §II. We then categorize low-resource IE approaches$^1$ into Traditional (§III-A) and LLM-based Methods (§III-B). In view of which aspect is chiefly enhanced when training, we further categorize traditional low-resource IE approaches into three paradigms:

1. Exploiting Higher-Resource Data (§III-A1): Increasing the size and diversity of the original sparse data with auxiliary resources which are generated endogenously or imported exogenously, aiming to obtain enriched samples and more precise semantic representations.
2. Developing Stronger Data-Efficient Models (§III-A2): Developing more robust models to better cope with maldistribution of samples and new unseen classes, aiming to improve model learning abilities so as to reduce dependence on samples.
3. Optimizing Data & Models Together (§III-A3): Jointly utilizing representative samples and robust models, aiming at adapting to low-resource scenarios promptly, and searching more suitable strategies for learning with sparse data.

We also divide LLM-based low-resource IE approaches based on whether the LLM is tuned, into two paradigms:

1. Direct Inference Without Tuning (§III-B1): Prompting LLMs to generate answers with given instructions, where prompts can either be in textual or code format.
2. Model Specialization With Tuning (§III-B2): Fine-tuning LLMs, either via prompt-tuning or standard fine-tuning. The fine-tuned LLMs are often tailored for specific IE tasks.

3 Due to page limits, complete papers are listed: https://github.com/zjunlp/Low-resource-KEPapers.
We then conduct empirical studies comparing traditional and LLM-based low-resource IE approaches in §IV. We explore the effectiveness of different techniques, especially those based on LLMs, to identify the most suitable approach for various low-resource IE tasks. The key findings are: (1) Well-tuned LMs continue to dominate in performance; (2) Tuning open-resource LLMs and using in-context learning (ICL) with GPTs presents promising results; (3) LLMs struggle to address the complex task with intricate schema. Besides, we summarize widely used benchmarks and promising applications in §V. Finally, we compare existing low-resource IE methods, discuss future directions in §VI, and make a conclusion in §VII.

II. PRELIMINARY ON LOW-RESOURCE IE

A. Information Extraction

Generally, IE [19], [20] can be regarded as structured prediction tasks [21], where a classifier is trained to approximate a target function \( F(x) \rightarrow y \), where \( x \in \mathcal{X} \) denotes the input data and \( y \in \mathcal{Y} \) denotes the output class sequence. This includes the subtasks of Named Entity Recognition (NER), Relation Extraction (RE) and Event Extraction (EE). For instance, given a sentence “Jack is married to the Iraqi microbiologist known as Dr. Germ.”:

- **Named Entity Recognition** [7] should identify the types of entities in given texts, e.g., ‘Jack’, ‘Dr. Germ’ ⇒ PERSON;
- **Relation Extraction** [8] should identify the relationship of the given entity pair (Jack, Dr. Germ) as husband_of;
- **Event Extraction** [9] should identify the event type as Marry, where the word ‘married’ triggers the event (subtask: ED, Event Detection); Jack and Dr. Germ are participants as husband and wife in the event respectively (subtask: EAE, Event Argument Extraction).

B. Low-Resource Scenarios

Most traditional IE models [22]–[24] assume that sufficient training data are indispensable to achieve satisfactory performance. However, in real-world applications, task-specific labeled samples tend to be unevenly distributed and new unseen classes may emerge over time, which result in low-resource scenarios [17]. Considering maldistribution of samples and new unseen classes, we systematically categorize low-resource scenarios into three aspects.

- **Long-tail Scenario** [25] means that only a minority of the classes are data-rich, while the majority of the classes have extremely few labeled sample data. Formally, given classes \( \mathcal{Y} = \mathcal{Y}_h \cup \mathcal{Y}_l \) with head classes \( \mathcal{Y}_h \) and long-tail classes \( \mathcal{Y}_l \), we denote the number of labeled samples for \( \mathcal{Y}_h \) and \( \mathcal{Y}_l \) as \( |\mathcal{Y}_h| \) and \( |\mathcal{Y}_l| \) respectively. \( |\mathcal{Y}_h| \) is much larger than \( |\mathcal{Y}_l| \), i.e., \( |\mathcal{Y}_h| \gg |\mathcal{Y}_l| \), while \( |\mathcal{Y}_h| \ll |\mathcal{Y}_l| \).

- **Few-shot Scenario** [26], [27] means that the classes for testing \( \mathcal{Y}_{te} \) have only a small number of samples, where the small number can be fixed following the N-Way-K-Shot setting [28] (each of the N classes contains only K samples), or unfixed and relatively small comparing with the total data size. Besides, \( \mathcal{Y}_{te} \) is nonexistent (unseen) in the training dataset with classes \( \mathcal{Y}_{tr} \), i.e., \( \mathcal{Y}_{tr} \cap \mathcal{Y}_{te} = \emptyset \).

- **Zero-shot Scenario** [29] means that the testing classes \( \mathcal{Y}_{te} \) to predict have no samples during the training phase. The model can classify these new unseen \( \mathcal{Y}_{te} \) by utilizing prior knowledge or relationships between seen and unseen classes, represented as class attributes or embeddings. In standard zero-shot scenarios, \( \mathcal{Y}_{te} \) only contains unseen classes in the training set; in generalized zero-shot scenarios [30], \( \mathcal{Y}_{te} \) contains both seen and unseen classes.

III. TAXONOMY OF TECHNICAL SOLUTIONS

In general, for this survey, we select influential low-resource IE papers mostly published within the last five years. The majority of these papers are sourced from prestigious conferences and journals, especially in NLP and machine learning domains, e.g., ACL, EMNLP, NAACL, NeurIPS, ICLR, TACL, TKDE and so on. We generally categorize existing approaches into traditional and LLM-based methods based on the model parameter size with a threshold: 6B. For Traditional Methods, the backbone model parameter size is less than 6B, e.g., BERT [174], BART [175], T5-large [176], GPT-2 [177], GPT-3\(^3\), etc. For LLM-based Methods, the backbone model parameter size is larger than 6B, e.g., Flan-T5-11B [178], LLaMA [179], GLM [180], GPT-3 [181], ChatGPT\(^3\), GPT-4 [182], etc. We refine representative methods in the two categories, and establish a fine-grained taxonomy as presented in Figure 1.

A. Traditional Methods

In view of which aspect is chiefly enhanced when addressing low-resource IE problems, we categorize some representative traditional methods into three general paradigms, as shown in Figure 1. In summary, traditional low-resource IE models tend to (1) exploit higher-resource data; (2) develop stronger data-efficient models; and (3) optimize data & models together.

1) Exploiting Higher-Resource Data: This paradigm mainly refers to data augmentation [183] or knowledge enhancement on the original small dataset, with endogenous or exogenous auxiliary resources. The goal is to create more representative samples and improve semantic representations with higher-resource data.

Weakly Supervised Augmentation involves synthesizing more training data through weak/distant supervision. This kind of methods usually utilize a knowledge base (KB) and some heuristic rules to automatically relabel training instances in corpus, potentially resulting in a noisy synthesized dataset.

[184] proposed distant supervision for RE, utilizing a large semantic KB — Freebase to label relations in an unlabeled corpus. Similarly, [31] used a dictionary for distantly supervised NER. [32] leveraged Freebase and FrameNet (an event KB) to automatically label training data for EE. Recent studies have proposed to improve quality of weakly supervised data with pretrained LMs (PLMs) [33], [34] and optimization strategies [35]–[39]. To mitigate selection bias, [57] have proposed a gradient imitation reinforcement learning framework for low-resource RE.

\(^1\)https://huggingface.co/EleutherAI/gpt-j-6b.
\(^3\)https://openai.com/blog/chatgpt.
**Multi-Modal Augmentation** supplements single modal with multi-modal samples to enhance semantics and facilitate disambiguation. Intuitively, the main challenge of such methods lies in effectively fusing data from different modalities. To fuse multi-modal data in low-resource IE, [40], [41] utilized attention mechanisms, [42]–[44] imposed multi-modal embedding space mapping, and [45], [46] leveraged prefix-guided multi-modal fusion.

**Multi-Lingual Augmentation** involves incorporating multi-lingual samples to achieve diverse and robust sample representation. Intuitively, the main challenge of such methods is to obtain language representations cross linguistics.

To tackle the issue, transferring cross-lingual knowledge [47]–[49] and capturing the consistency cross languages [50]–[52], [185] demonstrated effectiveness. Other promising models are to import additional context [54] and task-specific knowledge [53] for cross-lingual IE.

**Auxiliary Knowledge Enhancement** employs external knowledge as remedies, intended to learn semantic representation of samples more precisely. Different from weakly supervised augmentation using KB, this paradigm adopts more diverse knowledge formats and knowledge enhancement methods. Given diverse formats of auxiliary knowledge, we divide them into two categories.

(i) **Textual Knowledge**

Textual knowledge can be class-related knowledge [186], [187], like class descriptions [56], [188], [189] and class-specific texts [55]. It can also be synthesized data [58]–[63], [190]–[194] via data augmentation.

(ii) **Structured Knowledge**

Structured knowledge for low-resource IE can take the form of KG triples [195], task-specific ontology [196], and rules [65]. For example, [64] studied zero-shot RE by leveraging KG embeddings and logic rules to connect seen and unseen relations; [66] proposed a KB-aware NER framework to utilize class-heterogeneous knowledge in KBs; [67]–[69] resolved low-resource EE problems by utilizing association knowledge among classes.

2) Developing **Stronger Data-Efficient Models**: The paradigm focuses on developing models to handle sample...
maldistribution and unseen classes more effectively. The stronger models aim to improve learning abilities, so as to maximize the utilization of small data and minimize the dependence of samples.

**Meta Learning** [197] promptly assimilates emerging knowledge and deduces new classes by learning from few instances, with the ability of “learning to learn”, which is naturally suitable for few-shot IE tasks.

[70]–[79] utilized metric-based methods, mostly equipped with prototypical networks [28], [80]–[82], [198]–[200] leveraged model-agnostic methods based on MAML [201]. [83] proposed a memory-based method [202]. [84], [85] used model-based methods with Bayesian meta learning [203].

**Transfer Learning** [204] reduces the dependence on labeled target data by transferring learned class-invariant features, especially from high-resource to low-resource classes.

[86] leveraged class structures to transfer knowledge from existing to unseen classes. [87] proposed a weighted adversarial network to adapt features learned from high-resource to low-resource classes. [87], [91] utilized graph neural networks to facilitate knowledge transfer. [88], [92] used a memory module for information retrieval or similarity comparison from the source to target domain. [89] incorporated linguistic representations for cross-domain few-shot transfer.

**Fine-Tuning PLM** leverages PLMs [205] to utilize contextual representations and pre-trained parameters for fine-tuning. It adapts PLMs’ powerful language understanding capabilities to specific low-resource IE tasks.

Fine-tuning PLMs for low-resource IE aims at learning task-specific entity representations [96], [97], [104]; relation representations [93], [98], [102]; and event representations [94], [95], [99]–[101], [103] with data-efficient learning.

3) Optimizing Data & Models Together: This paradigm refers to jointly optimizing representative samples and data-efficient models, enabling swift adaptation to low-resource scenarios. The aim is to identify more suitable strategies for learning with sparse data.

**Multi-Task Learning** signifies learning multiple related tasks simultaneously by exploiting both task-generic commonality and task-specific diversity [169], contributing to improved performance for task-specific models, which naturally enables boost the target low-resource IE task.

(i) IE & IE-Related Tasks

[105], [206], [207] proposed to jointly model NER and Named Entity Normalization (NEN), as these two tasks can benefit each other with enhanced entity mention features. [106], [107] proposed to transfer the knowledge learned on Word Sense Disambiguation (WSD) to ED (a subtask of EE), considering ED and WSD are two similar tasks in which they both involve identifying the classes (i.e., event types or word senses) of some words in a given sentence.

(ii) Joint IE & OtherStructured Prediction Tasks

Different IE and structured prediction [21] tasks can also benefit from each other considering the similar task structures and progressive task process. For example, [108], [109] tackled the joint NER and RE task, respectively leveraging the relational graph and a copy mechanism; [110] incorporated global context in a general multi-task IE framework, w.r.t. the NER, RE and EE; [111], [112] jointly addressed IE and other structured prediction tasks, such as event-relation extraction, multi-span extraction, and n-ary tuple extraction.

**Task Reformulation** refers to formulating IE tasks into other formats which imports task-related knowledge, to capitalize on model architecture and data advantages. For example, IE can be reformulated as machine reading comprehension (MRC) or text-to-structure generation tasks. MRC-based IE identifies answer spans in context of questions, providing crucial knowledge for target tasks. Generative IE [208]–[210] employs generative LMs (GenLMs) for IE, which can reduce error propagation while increasing adaptability for IE.

(i) Reformulating IE as QAMRC. [115] tackled low-resource EE by transferring event schema into natural questions. [116], [117], [211] studied how question generation strategies affect QA-based EE. [113], [114], [212] explored QA-based IE to efficiently encode of crucial information of classes.

(ii) Reformulating IE as Text-to-Structure Generation. [119] proposed a sequence-to-structure generation paradigm for EE. [120] framed IE as a translation task, efficiently extracting task-relevant information. [121] employed sequence-to-structure for zero-shot RE. [122], [123]–[128], [214] respectively applied a template-based, template-free, dynamic-template-filling, demonstration-based, lightweight and contrastive prompt-based method for few-shot NER. [125], [129] leveraged the structured template and discriminative soft prompts for zero-shot RE. [124], [130], [215], [216] respectively utilized template-based, extractive, contextualized, and type-specific prompts for low-resource EE.

**Prompt-Tuning PLM** [213] involves inserting text pieces, i.e., templates, into the input to convert a classification task into a masked language modeling problem. This enables IE approaches to benefit from LMs’ pretrained knowledge, improving sample efficiency.

(i) Vanilla Prompt Tuning

Vanilla prompt tuning methods leverage the basic prompt learning framework, excelling in low-resource scenarios. [122], [123], [126]–[128], [214] respectively applied a template-based, template-free, dynamic-template-filling, demonstration-based, lightweight and contrastive prompt-based method for few-shot NER. [125], [129] leveraged the structured template and discriminative soft prompts for zero-shot RE. [124], [130], [215], [216] respectively utilized template-based, extractive, contextualized, and type-specific prompts for low-resource EE.

(ii) Augmented Prompt Tuning

Augmented prompt tuning methods enhance vanilla prompt learning by incorporating diverse knowledge, facilitating low-resource IE. [131] incorporated knowledge among labels to prompt-tuning. [134], [141] enhanced prompts with label semantics and class description. [15], [133], [135], [140] demonstrated the effectiveness of enhancing prompt-tuning with ontology, rules, semantic structures and reasoning rationales. [136], [138] introduced a unified text generation (UIE) and semantic matching (USM) framework for different IE tasks. [137] proposed a structure-aware GenLM to harness syntactic knowledge for UIE. [139] introduced UIE for any kind of schemas. [132], [217]–[219] utilized retrieval-augmented prompts to import task-specific knowledge.
B. LLM-Based Methods

Comparing with traditional PLMs, LLMs possess more powerful pretrained abilities, allowing for more complex prompt learning. In view of whether the LLM is tuned (i.e., LLM initial parameters are modified), we categorize some representative LLM-based methods into two general paradigms, as shown in Figure 1.

1) Direct Inference without Tuning: This kind of methods typically involve ways to leverage LLMs without extensive additional training. They can boost low-resource IE by leveraging the inherent capabilities of LLMs to understand and process contexts, further obtaining valuable insights from scarce data, thereby reducing the requirement of fine-tuning.

Instruction Prompting involves giving the LLM explicit instructions (without demonstrations) to perform a specific task. For low-resource IE, instruction prompting can be effective as it allows the model to execute tasks using its pre-existing knowledge and language understanding.

As instruction prompting does not require demonstrations, it is naturally suitable for zero-shot [143], [147], [148], [220] and cross-domain [146] IE tasks. [142] investigated zero-shot ED task and found that ChatGPT is competitive in simple scenarios, but struggles in more complex and long-tail scenarios. [144] used global constraints with prompting for zero-shot EE, demonstrating adaptability to any other datasets. [145], [221] observed that ChatGPT is advantageous in zero-shot NER and RE tasks with instruction prompting.

Code Prompting involves presenting the LLM with snippets of code (or code-like instructions [154]) to guide its responses. This kind of methods can be particularly useful in low-resource IE tasks that involve structured output, as the code implies the schema of the specific task.

[149]–[151] tackled low-resource IE tasks with code prompting on Code-LLMs, such as CodeX [222], demonstrating the effectiveness of code-style prompts. [153] proposed a universal retrieval-augmented code generation framework. Moreover, code prompting can also be applied to multimodal IE tasks [152].

In-Context Learning (ICL) utilizes the ability of LLMs to learn from the context provided in prompts. The model uses few relevant examples (demonstrations) to "understand" the specific IE task and then applies this understanding to new data, particularly useful in low-resource scenarios.

Recent research leveraged ICL for low-resource NER [155], [160], [162], [167], [168], RE [156], [157], [161], [223], joint IE [158], [159], [163], [164] and OpenIE [165], [166]. The key challenges of ICL for IE are: (1) input prompting fails to thoroughly express intricate IE tasks, and (2) aligning input and labels is not effective enough. To tackle these issues, synthesizing data with LLMs [224], [225] and inputting more task-specific prompts are promising.

2) Model Specialization with Tuning: This kind of methods enhance low-resource IE by tailoring the model capabilities to specific tasks, and can be divided into prompt-tuning and fine-tuning. Generally, prompt-tuning is particularly valuable for its efficiency and minimal data requirements. Fine-tuning, while more resource-intensive, offers deeper customization and potentially better performance.

Prompt-Tuning LLM involves keeping the LLM weights fixed and only tuning a small set of parameters associated with the prompts. In low-resource IE tasks, prompt-tuning allows for adapting the model to specific tasks or domains with minimal data, where prompts act as a guide.

[173] introduced a benchmark of diverse IE tasks with expert-written instructions, and proposed a unified IE framework, InstructUIE, with instruction tuning on FlanT5-11B [178]. [171] proposed InstructIE dataset and fine-tune LLama-7B [179] with instruction-following capability on IE. [170] hypothesized that instruction-tuning has been unable to elicit strong RE capabilities in LLMs by aligning RE with QA. [169] pretrained GLM-10B [180] on a collection of task-agnostic corpora to generate structures from text. [172] explored targeted distillation with mission-focused instruction tuning to train student models on NER task.

Fine-Tuning LLM involves adjusting the weights of the LLM on a smaller and task-specific dataset. This approach is more data-intensive than prompt-tuning but expectantly leads to performance gains.

Actually, this kind of methods are currently underdeveloped for low-resource IE tasks due to the limit of sufficient computing resources.

IV. EMPIRICAL STUDY ON TECHNICAL SOLUTIONS

A. Setup

Datasets. We adopt some widely used datasets for each subtask of low-resource IE. NER: we use CoNLL03 [227]; OntoNotes5.0 [228]; FewNERD [229]. RE: we adopt TACREVF [230]; NYT [231]; FewRel [232]. ED of EE: we use ACE05 [233]; MAVEN [234]; FewEvent [71]. EAE (event argument extraction) of EE: we adopt ACE05 [233]; RAMS [235]; WikiEvents [236].

We follow [159] to construct few-shot datasets in this paper. In all experiments, we utilize micro F1 score following previous methods like [136] to evaluate performance. We release the code and data on GitHub.4

Models. We conduct empirical study on LLM-based methods comparing with previous SOTA approaches. For prompt/fine-tuning LLM, we adopt three typical models: InstructUIE [173] with prompt-tuning on FlanT5-11B [178]; KnowLM5 which fine-tuned LLama [179] with LoRA [237]; and fine-tuning ChatGPT6. For ICL with LLM, we select three widely used LLMs: Vicuna(1.5)7; ChatGPT3; GPT-4 [182].

B. Implementation Details

We test InstructUIE and KnowLM following their guidelines. In addition to the default parameters, for InstructUIE, we respectively set the maximum length for source, target and generation to 512, 50, 50, and set maximum instance

4https://github.com/mayubo2333/LLM_project.
5https://github.com/zjunlp/KnowLM.
6https://platform.openai.com/docs/guides/fine-tuning.
7https://github.com/ml-sys/FastChat.
number for each task to 200; for KnowLM, we use the version of knowlm-13b-base-v1.0, and respectively set the maximum length for source and new tokens to 512 and 300. For Vicuna, we adopt Vicuna-13B and Vicuna-1.5-13B without fine-tuning, setting the maximum input length to 1800; the batch size to 1; and both frequency_penalty and presence_penalty to 0. About the number of demonstrations, we set it to 4 for NER and ED; 8 for EAE; and 16 for RE. We run each experiment on 4 NVIDIA V100 GPUs for InstructUIE and KnowLM, and a single one for Vicuna. To save memory, we leverage the Accelerate framework and fp16 inference. For Fine-Tuning ChatGPT, we fine-tune gpt-turbo-3.5-0613 on each specific dataset. We split 10% samples from training set to construct valid set and train over 5 epochs with 5-shot samples, then conduct instruction prompting on fine-tuned ChatGPT. For ChatGPT and GPT-4, we call official APIs (gpt-3.5-turbo-0301/gpt-4-0314) without fine-tuning, and set the maximum input length to 3600 for all tasks. For ICL with LLMs (Vicuna1.5, ChatGPT, GPT-4), we generate output with sampling temperature as 0 (i.e., greedy decoding). We unify the maximum output length to 32 for RE; 96 for NER, ED and EAE.

C. Experimental Analysis

In this section, we intend to analyze the performance of various (especially LLVM-based) models, and explore the optimal selection of low-resource IE approaches. We summarize main experimental results in Table I, and elaborate on our key findings as below:

### Open-source or Proprietary LLMs?

We observe that prompt-/fine-tuning InstructUIE/KnowLM (open-source LLMs) significantly outperform fine-tuned ChatGPT (proprietary LLMs) on most tasks, except for under-trained datasets, e.g., TACREv, MAVEN, RAMS and WikiEvents. Nevertheless, ICL with Vicuna1.5 (open-source LLMs) lag far behind ChatGPT and GPT-4 (proprietary LLMs). Thus we conclude that for ICL, proprietary LLMs outperform open-source LLMs; when it comes to prompt-tuning or fine-tuning, open-source LLMs outperform their proprietary counterparts. We speculate that the possible reasons are: (1) The fundamental abilities of proprietary LLMs are much better (larger parameters, more abundant training data and effective training strategies), so that they understand the task from in-context demonstrations more effectively; (2) Open-source LLMs are much more lightweight than proprietary LLMs, contributing to thoroughly training and adapting to a specific task. While for proprietary LLMs like ChatGPT, the limited flexibility of fine-tuning APIs makes it difficult to obtain a well-trained task-specific model.

### Tuning or ICL?

Regarding tuning InstructUIE and KnowLM, we find they achieve satisfying performance on well-trained tasks (e.g., NER and RE) but show limited generalization abilities on under-trained tasks (e.g., ED and EAE). For ChatGPT, we observe that fine-tuning leads to generally consistent improvement on RE, ED and EAE tasks than ICL, especially on datasets which ICL with ChatGPT struggles, e.g., NYT, TACREv, ACE05 and MAVEN. However, even fine-tuned ChatGPT only shows similar performance with GPT-4 and performs worse than well-tuned SOTA small PLMs and open-resource LLMs. The possible reason is that for ICL with not-fine-tuned LLMs, where the instructions of LLMs are under alignment with the specific IE task [170], it could be difficult to fully leverage the power of LLMs. As prompt-/fine-tuning enables adapting the LLM to specific tasks by adjusting its inputs or training it on specific data, we can infer that tuning performs generally best in low-resource IE, and more thoroughly training contributes to better performance.

### Simple v.s. Complex Tasks

Across different IE tasks, we observe that LLMs are more proficient on NER and EAE (event class given) tasks than ED task. Given that ED is a more complex task (which requires instantiating abstract event classes on words in the context), we speculate that prompt-tuned LLMs are not well-acquainted with this task, and unable to fully understand the complex task through instructions and demonstrations. Furthermore, we find that the performance...
Table II: A SUMMARY OF SOME PUBLICLY-RELEASED LOW-RESOURCE IE BENCHMARKS.

| Task       | Dataset                  | Summary                                                                 |
|------------|--------------------------|-------------------------------------------------------------------------|
| Low-res NER| Few-NERD [229]           | human-annotated; 8 super- and 66 sub-entity types; 188,238 sentences from Wikipedia with 4,601,160 words. |
| Low-res RE | FewRel [232]             | human-annotated; 100 relations; 70,000 sentences derived from Wikipedia with 124,577 unique tokens in total. |
|            | FewRel2.0 [238]          | more challenging than FewRel, considering biomedical domain and NOTA setting. |
| Low-res EE | Entail-RE [69]           | human-annotated; 80 relation types; 220k instances from Wikipedia; 19 relation entailment pairs. |
|            | FewEvent [71]            | human- & machine-annotated; 19 super- and 100 sub-event types; 70,852 instances from Wikipedia. |
|            | Causal-EE [69]           | human-annotated; 80 event types; 3.7k instances from Wikipedia; 112 event causality pairs. |
|            | OntoEvent [68]           | human- & machine-annotated; 13 super- and 100 sub-event types; 60,546 instances from Wikipedia; 3,804 event correlation pairs. |

A large-scale dataset is proposed for zero-shot RE. More low-resource RE studies are shown in FewREBench [239].

Low-resource EE. FewEvent [71] designed for few-shot ED, is derived from widely-used EE datasets and data augmentation. Causal-EE [69] and OntoEvent [68], designed for low-resource ED, is augmented with event-relations. FewDocAE [240] is used for few-shot document-level EAE. More EE studies are shown in TextEE [241].

V. BENCHMARKS AND APPLICATIONS

A. Benchmarks

In this section, we briefly introduce some public benchmarks specially established for low-resource IE, as shown in Table II. Besides, there are also some low-resource IE methods [115] directly sampling part of data from public IE benchmarks in general scenarios.

Low-resource NER. Few-NERD [229] is designed for few-shot NER, where each fine-grained entity type has sufficient examples for few-shot learning, including Few-NERD (INTRA) and Few-NERD (INTER). All entities in different sets of Few-NERD (INTRA) belong to different types, in contrast with Few-NERD (INTER).

Low-resource RE. FewRel [232], designed for few-shot RE, is a large-scale supervised dataset. FewRel2.0 [238] is derived from FewRel, and used for more challenging tasks: (1) adapt to a new domain with only a handful of instances, and (2) detect none-of-the-above (NOTA) relations. Entail-RE [69], designed for low-resource RE, is augmented with relation entailment annotations. Wiki-ZSL [56] is proposed for zero-shot RE. More low-resource RE studies are shown in FewREBench [239].

Low-resource EE. FewEvent [71] designed for few-shot ED, is derived from widely-used EE datasets and data augmentation. Causal-EE [69] and OntoEvent [68], designed for low-resource ED, is augmented with event-relations. FewDocAE [240] is used for few-shot document-level EAE. More EE studies are shown in TextEE [241].

B. Applications

We propose some promising applications of low-resource IE, and summarize them into two aspects.

1) Domain-specific Applications [3]. With the continuous emergence of knowledge from various domains, such as health care, natural science, and finance, it’s challenging to annotate large-scale structured knowledge. Especially for specific domains involving data privacy, where sufficient labeled data is not available, low-resource IE is naturally crucial.

2) Knowledge-intensive Tasks [5]. Low-resource IE has the potential to enhance knowledge-intensive tasks such as KG-based QA, fact verification, text-to-structure generation, and commonsense reasoning. Intuitively, those tasks require background KG for reasoning; however, collecting high-quality KGs is cumbersome and time-consuming. Low-resource IE can efficiently collect diverse knowledge across resources and simultaneously enhance performance.

VI. COMPARISON, DISCUSSION AND OUTLOOK

Comparison of Traditional Methods. From the methodology perspective, (1) Exploiting higher-resource data, essentially to enrich the data features in low-resource scenarios, serves as a general solution, which can be applied to widespread backbone models. (2) Developing stronger data-efficient models, essentially to constrain the hypothesis space more precisely, is a fundamental advancement to enable models equipped with specific low-resource learning abilities, typically through leveraging or transferring “modeledge” (e.g., implicit knowledge captured in model weights) in PLMs. (3) Optimizing data & models together, essentially to pursue the best strategy for training on representative data and searching the optimal hypothesis space, represents a synthesis of the aforementioned two paradigms.

From the task perspective, low-resource NER and RE are two typical tasks that have been well investigated. Since their schema of NER and RE are relatively simple, they can normally obtain satisfactory performance, especially with the utilization of traditional PLMs and recent LLMs. However, low-resource EE still suffers from poor generalization due to the complicated task schema. Thus, it remains challenging but promising to resolve the EE issues of event interpretation and representation in low-resource scenarios.

Discussion on LLM-Based Methods. Comparing to traditional PLMs, LLMs equipped with more powerful pretrained abilities enable more complex prompt learning, and can be given more complicated instructions. Similar to traditional PLMs, LLMs with limited input capacity also struggle to
tackle IE tasks with the intricate schema. Through the empirical study in §IV-C, we deduce that tuning open-resource LLMs and ICL with GPT family is promising in general, and the optimal LLM-based technical solution for low-resource IE can be task-dependent.

Furthermore, [159] demonstrated that LLMs are generally not effective few-shot information extractors, but excel at reranking challenging samples. Besides, LLMs have potential to serve as data creators [242] for low-resource IE. In addition, as LLMs allow for multi-agent interaction [243], [244], tool use [221], [245], etc, it’s promising to achieve low-resource IE through collaboration between different models, e.g., large and small LM cooperation.

Future Directions. Despite many models were proposed as surveyed, potential directions remain:

- Utilizing more informative knowledge [246]–[249]. Advanced sample selection techniques, like active learning, sample reranking [159], [250] and data synthesizing [225], [242] with LLMs, are promising to efficiently identify and utilize informative samples, reducing large-scale annotation requirements.

- Focusing on practical low-resource settings and applications. Current low-resource IE models are mostly research-oriented which is unrealistic in real world settings. It is essential to explore practical low-resource IE applications, especially involving out-of-distribution (OOD) data, establishing benchmarks and evaluation metrics (e.g., low-computation/memory cost).

- Equipping models with adaptable inference. We assume that low-resource IE should be robust to domain shifts, rather than being restricted to a specific domain. Thus, exploring lifelong low-resource IE and OpenIE [165], [166], [251] then schema induction [252] are promising.

- Exploring robust, faithful and interpretable IE. To help LMs exactly understand IE targets and rationales [15] can also be effective, especially with utilization of LLMs. Additionally, incorporating human feedback in training can further enhance accuracy, fairness, and reduce risks of biased or discriminatory IE.

VII. CONCLUSION

In this paper, we present a literature review on low-resource IE methodologies from traditional and LLM-based perspectives, systematically categorizing traditional methods into three paradigms: (1) exploiting higher-resource data, (2) developing stronger data-efficient models, and (3) optimizing data and models together; and LLM-based methods into two paradigms: (1) direct inference without tuning, and (2) model specialization with tuning. We also summarize widely used benchmarks and suggest some applications. Furthermore, we compare traditional paradigms, discuss low-resource IE with LLMs, and provide some insights into future work. Our survey aims to assist researchers in understanding this field and inspire innovative algorithms, while guiding industry practitioners in selecting appropriate technical solutions for real-world applications.

ACKNOWLEDGMENT

We would like to express gratitude to the anonymous reviewers for their kind and helpful comments. This work was supported by the National Natural Science Foundation of China (No. 62206246), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), Tencent AI Lab Rhino-Bird Focused Research Program (RBF2024003), Information Technology Center and State Key Lab of CAD&CG, Zhejiang University, and NUS-NCS Joint Laboratory (A-0008542-00-00).

REFERENCES

[1] J. R. Cowie and W. G. Lehnert, “Information extraction,” Commun. ACM, vol. 39, no. 1, pp. 80–91, 1996. [Online]. Available: https://doi.org/10.1145/234173.234209
[2] R. Grishman, “Information extraction,” IEEE Intell. Syst., vol. 30, no. 5, pp. 8–15, 2015. [Online]. Available: https://doi.org/10.1109/MIS.2015.68
[3] Y. Wang, L. Wang, M. Rastegar-Mojarrad, S. Moon, E. Shen, N. Goual, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, “Clinical information extraction applications: A literature review,” J. Biomed. Informatics, vol. 77, pp. 34–49, 2018. [Online]. Available: https://doi.org/10.1016/j.jbi.2017.11.011
[4] S. Deng, C. Wang, Z. Li, N. Zhang, Z. Dai, H. Chen, F. Xiong, M. Yan, Q. Chen, M. Chen, J. Chen, J. Z. Pan, B. Hooi, and H. Chen, “Construction and applications of billion-scale pre-trained multimodal business knowledge graph,” in ICDE. IEEE, 2023, pp. 2988–3002. [Online]. Available: https://doi.org/10.1109/ICDE55515.2023.00229
[5] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, “A survey on knowledge graphs: Representation, acquisition, and applications,” IEEE Trans. Neural Networks Learn. Syst., vol. 33, no. 2, pp. 494–514, 2022. [Online]. Available: https://doi.org/10.1109/TNNLS.2021.3070843
[6] S. Deng, N. Zhang, W. Zhang, J. Chen, J. Z. Pan, and H. Chen, “Knowledge-driven stock trend prediction and explanation via temporal convolutional network,” in WWW (Companion Volume). ACM, 2019, pp. 678–685. [Online]. Available: https://doi.org/10.1145/3308560.3317701
[7] J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 1, pp. 50–70, 2022. [Online]. Available: https://doi.org/10.1109/TKDE.2020.2981314
[8] H. Wang, K. Qin, R. Y. Zakari, G. Lu, and J. Yin, “Deep neural network-based relation extraction: an overview,” Neural Comput. Appl., vol. 34, no. 6, pp. 4781–4801, 2022. [Online]. Available: https://doi.org/10.1007/s00521-021-01666-3
[9] Q. Li, J. Li, J. Sheng, S. Cui, J. Wu, Y. Hei, H. Peng, S. Guo, L. Wang, A. Beheshti, and P. S. Yu, “A survey on deep learning event extraction: Approaches and applications,” IEEE Trans. Neural Networks Learn. Syst., pp. 1–21, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9927311
[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. Wen, “A survey of large language models,” CoRR, abs/2303.18223, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.18223
[11] Z. Zhang, H. Li, Z. Zhang, Y. Qin, X. Wang, and W. Zhu, “Large graph models: A perspective,” CoRR, vol.abs/2303.14522, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.14522
[12] J. Huang, C. Li, K. Sabubdi, D. Jose, S. Balakrishnan, W. Chen, B. Peng, J. Gao, and J. Han, “Few-shot named entity recognition: An empirical baseline study,” in EMNLP (1). Association for Computational Linguistics, 2021, pp. 10 408–10 423. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.813
[13] V. Moscato, M. Postiglione, and G. Sperl, “Few-shot named entity recognition: Definition, taxonomy and research directions,” ACM Trans. Intell. Syst. Technol., vol. 14, no. 5, pp. 94:1–94:46, 2023. [Online]. Available: https://doi.org/10.1145/3609483

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, “Exploring pre-trained J. Fang, X. Wang, Z. Meng, P. Xie, F. Huang, and Y. Jiang, N. Zhang, S. Deng, Z. Sun, J. Chen, W. Zhang, and H. Chen, T. Zhang, C. Xia, C. Lu, and P. S. Yu, “MZET: memory augmented model for cross long.414 M. Qu, T. Gao, L. A. C. Xhonneux, and J. Tang, “Few-shot relation extraction via bayesian meta-learning on relation graphs,” in ICML, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 7867–7876. [Online]. Available: http://proceedings.mlr.press/v119/qual20a.html S. Shen, T. Wu, G. Qi, Y. Li, G. Haffari, and S. Bi, “Adaptive knowledge-enhanced bayesian meta-learning for few-shot event detection,” in ACL/IJCNLP (Findings). Association for Computational Linguistics, 2021, pp. 2417–2429. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.214 L. Huang, H. Ji, K. Cho, I. Dagan, S. Riedel, and C. R. Voss, “Zero-shot transfer learning for event extraction,” in ACL (1). Association for Computational Linguistics, 2018, pp. 2160–2170. [Online]. Available: https://www.aclweb.org/anthology/A18-1165/ N. Zhang, S. Deng, Z. Sun, G. Wang, X. Chen, W. Zhang, and H. Chen, “Long-tail relation extraction via knowledge graph embeddings and graph convolution networks,” in NAACL-HLT (1). Association for Computational Linguistics, 2019, pp. 3016–3025. [Online]. Available: https://doi.org/10.18653/v1/n19-1306 T. Zhang, C. Xia, C. Lu, and P. S. Yu, “MZET: memory augmented zero-shot fine-grained named entity typing,” in COLING. International Committee on Computational Linguistics, 2020, pp. 77–87. [Online]. Available: https://doi.org/10.18653/v1/2020.coling-main.7 S. Gururaja, R. Dutt, T. Liao, and C. P. Rosé, “Linguistic representations for fewer-shot relation extraction across domains,” in ACL (1). Association for Computational Linguistics, 2023, pp. 7502–7514. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.414 N. Zhang, S. Deng, Z. Sun, J. Chen, W. Zhang, and H. Chen, “Relation adversarial network for low resource knowledge graph completion,” in WWW. ACM, 2020, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3366423.3380089 J. Li, S. Feng, and B. Chiu, “Few-shot relation extraction with dual graph neural network interaction,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–13. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10143375 J. Fang, X. Wang, Z. Meng, P. Xie, F. Huang, and Y. Jiang, “MANNER: A variational memory-augmented model for cross domain few-shot named entity recognition,” in ACL (1). Association for Computational Linguistics, 2023, pp. 4261–4276. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.234 L. B. Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski, “Matching the blanks: Distributional similarity for relation learning,” in ACL (1). Association for Computational Linguistics, 2019, pp. 2905–2905. [Online]. Available: https://doi.org/10.18653/v1/p19-1279 S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, “Exploring pre-trained language models for event extraction and generation,” in ACL (1). Association for Computational Linguistics, 2019, pp. 5284–5294. [Online]. Available: https://doi.org/10.18653/v1/2019.findings-acl.16 S. Wang, M. Yu, S. Chang, L. Sun, and L. Huang, “Query and extract: Refining event extraction as type-oriented binary decomposing,” in ACL (Findings). Association for Computational Linguistics, 2022, pp. 169–182. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-acl.16 M. Xue, B. Yu, Z. Zhang, T. Liu, Y. Zhang, and B. Wang, “Coarse-to-fine pre-training for named entity recognition,” in EMNLP (1). Association for Computational Linguistics, 2020, pp. 6345–6354. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.514 J. Chen, Q. Liu, H. Lin, X. Han, and L. Sun, “Few-shot named entity recognition with self-describing networks,” in ACL (1). Association for Computational Linguistics, 2022, pp. 5711–5722. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.392 W. Zhou, S. Zhang, T. Naumann, M. Chen, and H. Poon, “Continual contrastive fine-tuning improves low-resource relation extraction,” in ACL (1). Association for Computational Linguistics, 2023, pp. 13 249–13 263. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.739 Z. Wang, X. Wang, X. Han, Y. Lin, L. Hou, Z. Liu, P. Li, J. Li, and J. Zhou, “CLEVE: contrastive pre-training for event extraction,” in ACL/IJCNLP (1). Association for Computational Linguistics, 2021, pp. 6283–6297. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.491 A. P. B. Veyseh, V. D. Lai, F. Dernoncourt, and T. H. Nguyen, “Unleash GPT-2 power for event extraction,” in ACL/IJCNLP (1). Association for Computational Linguistics, 2021, pp. 6271–6282. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.490 Y. Zhou, T. Shen, X. Geng, G. Long, and D. Jiang, “Claret: Pre-training a correlation-aware context-to-event transformer for event-centric named entity recognition,” in Findings of ACL/IJCNLP (1). Association for Computational Linguistics, 2022, pp. 2559–2575. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.183 S. S. S. Das, H. Zhang, P. Shi, W. Yin, and R. Zhang, “Unified low-resource sequence labeling by sample-aware dynamic sparse fine-tuning,” in EMNLP. Association for Computational Linguistics, 2022, pp. 6998–7010. [Online]. Available: https://aclanthology.org/2022.acl-IJCNLP.373 H. Zhang, W. Yao, and D. Yu, “Efficient zero-shot event extraction with context-definition alignment,” in EMNLP (Findings). Association for Computational Linguistics, 2022, pp. 7169–7179. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-emnlp.531 G. Dong, Z. Wang, J. Zhao, G. Zhao, D. Guo, D. Fu, T. Hui, C. Zeng, K. He, X. Li, L. Wang, X. Cui, and W. Xu, “A multi-task semantic decomposition framework with task-specific pre-training for few-shot NER,” in CIKM. ACM, 2023, pp. 430–440. [Online]. Available: https://doi.org/10.1145/3583780.3614766 H. Zhang, W. Yao, and D. Yu, “Efficient zero-shot event extraction with context-definition alignment,” in EMNLP (Findings). Association for Computational Linguistics, 2022, pp. 6998–7010. [Online]. Available: https://aclanthology.org/2022.acl-IJCNLP.373 B. Zhou, X. Cai, Y. Zhang, and X. Yuan, “An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization,” in ACL/IJCNLP (1). Association for Computational Linguistics, 2022, pp. 6214–6224. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-IJCNLP.485 W. Lu and T. H. Nguyen, “Similar but not the same - word sense disambiguation improves event detection via neural representation matching,” in EMNLP. Association for Computational Linguistics, 2018, pp. 4822–4828. [Online]. Available: https://aclanthology.org/D18-1517/ Y. D. Lai, M. V. Nguyen, T. H. Nguyen, and F. Dernoncourt, “Graph learning regularization and transfer learning for few-shot event detection,” in SIGIR. ACM, 2021, pp. 2172–2176. [Online]. Available: https://doi.org/10.1145/3404835.3463054 T. Fu, P. Li, and W. Ma, “Graphrl: Modeling text as relational graphs for joint entity and relation extraction,” in ACL (1). Association for Computational Linguistics, 2019, pp. 1409–1418. [Online]. Available: https://doi.org/10.18653/v1/p19-1136 D. Zeng, H. Zhang, and Q. Liu, “Copyml: Copy mechanism for joint extraction of entities and relations with multi-task learning,” in AAAI. AAAI Press, 2020, pp. 9507–9514. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6495 D. Wadden, U. Wernberg, Y. Luan, and H. Hajiishiriz, “Entity, relation, and event extraction with contextualized span representations,” in EMNLP/IJCNLP (1). Association for Computational Linguistics, 2019, pp. 5783–5788. [Online]. Available: https://doi.org/10.18653/v1/2019.acl-IJCNLP.495 S. Deng, S. Mao, N. Zhang, and B. Hooi, “Speech: Structured prediction with energy-based event-centric hyperspheres,” in ACL (1). Association for Computational Linguistics, 2023, pp. 351–363. [Online]. Available: https://aclanthology.org/2023.acl-long.21/
[236] S. Li, H. Ji, and J. Han, “Document-level event argument extraction by conditional generation,” in NAACL-HLT: Association for Computational Linguistics, 2021, pp. 984–988. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.69

[237] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in ICLR. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=SezVEKeeFY9

[238] T. Gao, X. Han, H. Zhu, Z. Liu, P. Li, M. Sun, and J. Zhou, “Fewrel 2.0: Towards more challenging few-shot relation classification,” in EMNLP/IJCNLP (1). Association for Computational Linguistics, 2019, pp. 6249–6254. [Online]. Available: https://doi.org/10.18653/v1/D19-1649

[239] X. Xu, X. Chen, N. Zhang, X. Xie, X. Chen, and H. Chen, “Towards realistic low-resource relation extraction: A benchmark with empirical baseline study,” in EMNLP (Findings), 2022, pp. 413–427. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.29

[240] X. Yang, Y. Lu, and L. R. Petzold, “Few-shot document-level event argument extraction,” in ACL (1). Association for Computational Linguistics, 2023, pp. 8029–8046. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.446

[241] K. Huang, I. Hsu, T. Parekh, Z. Xie, Z. Zhang, P. Natarajan, K. Chang, N. Peng, and H. Ji, “A reevaluation of event extraction: Past, present, and future challenges,” CoRR, vol. abs/2311.09562, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.09562

[242] D. Lee, J. Pujara, M. Sewak, R. White, and S. K. Jauhar, “Making large language models better tool learners with execution feedback,” CoRR, vol. abs/2309.17234, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.17234

[243] J. Zhang, X. Xu, and S. Deng, “Exploring collaboration mechanisms for LLM agents: A social psychology view,” CoRR, vol. abs/2310.02124, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.02124

[244] S. Qiao, H. Gui, H. Chen, and N. Zhang, “Making language models better tool learners with execution feedback,” CoRR, vol. abs/2305.13068, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.13068

[245] Z. Feng, W. Ma, W. Yu, L. Huang, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, “Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications,” CoRR, vol. abs/2311.05876, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.05876

[246] T. Pan, L. Luo, Y. Wang, C. Chen, and X. Wu, “Unifying large language models and knowledge graphs: A roadmap,” CoRR, vol. abs/2306.08302, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.08302

[247] J. Z. Pan, S. Razniewski, J. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omelyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, and D. Graux, “Large language models and knowledge graphs: Opportunities and challenges,” Transactions on Graph Data and Knowledge, vol. 1, no. 1, pp. 2:1–2:38, 2023. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/TGD.1.1.2

[248] H. Chen, “Large knowledge model: Perspectives and challenges,” 2023. [Online]. Available: https://arxiv.org/abs/2312.02706

[249] A. Amalvy, V. Labatut, and R. Dufour, “Learning to rank context for named entity recognition using a synthetic dataset,” in EMNLP. Association for Computational Linguistics, 2023, pp. 10372–10382. [Online]. Available: https://aclanthology.org/2023.emnlp-main.642

[250] A. Amalvy, V. Labatut, and R. Dufour, “Learning to rank context for named entity recognition using a synthetic dataset,” in EMNLP. Association for Computational Linguistics, 2023, pp. 10372–10382. [Online]. Available: https://aclanthology.org/2023.emnlp-main.642

[251] J. Qi, C. Zhang, X. Wang, K. Zeng, J. Yu, J. Liu, L. Hou, J. Li, and X. Bin, “Preserving knowledge invariance: Rethinking robustness evaluation of open information extraction,” in EMNLP: Association for Computational Linguistics, 2023, pp. 5876–5890. [Online]. Available: https://aclanthology.org/2023.emnlp-main.360

[252] T. Zhang, I. Tham, Z. Hou, J. Ren, L. Zhou, H. Xu, L. Zhang, L. J. Martin, R. Dror, S. Li, H. Ji, M. Palmer, S. W. Brown, R. Suchocki, and C. Callison-Burch, “Human-in-the-loop schema induction,” in ACL (demo). Association for Computational Linguistics, 2023, pp. 1–10. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-demo.1

[253] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, “Automated concatenation of embeddings for structured prediction,” in ACLI/JCNLP (1). Association for Computational Linguistics, 2021, pp. 2643–2660. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.206

[254] Y. Fei and X. Xu, “GFMRC: A machine reading comprehension model for named entity recognition,” Pattern Recognit. Lett., vol. 172, pp. 97–105, 2023. [Online]. Available: https://doi.org/10.1016/j.patrec.2023.06.011

[255] D. Ye, Y. Lin, P. Li, and M. Sun, “Packed lexemarker for entity and relation extraction,” in ACL (1). Association for Computational Linguistics, 2022, pp. 4904–4917. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.337

[256] Z. Yang, Y. Liu, and C. Ouyang, “Causal intervention-based few-shot named entity recognition,” in EMNLP (Findings). Association for Computational Linguistics, 2023, pp. 15635–15646. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.1046

[257] Y. Zhan, Z. Li, X. Zhao, C. Zhang, and T. Chen, “A simple overlapping relation extraction method based on dropout,” in IJCNN. IEEE, 2022, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN55064.2022.9892310

[258] J. Zheng and Z. Chen, “Sentence-level relation extraction via contrastive learning with descriptive relation prompts,” CoRR, vol. abs/2304.04935, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.04935

[259] X. Ma, X. Qin, J. Liu, and W. Ran, “Interaction information guided prototype representation rectification for few-shot relation extraction,” Electronics, vol. 12, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/13/2912

[260] S. Zhang, T. Ji, W. Ji, and X. Wang, “Zero-shot event detection based on ordered contrastive learning and prompt-based prediction,” in NAACL-HLT (Findings). Association for Computational Linguistics, 2022, pp. 2572–2580. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-naacl.196

[261] W. Liu, S. Cheng, D. Zeng, and H. Qu, “Enhancing document-level event argument extraction with contextual clues and role relevance,” in ACL (Findings). Association for Computational Linguistics, 2023, pp. 12908–12922. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-acl.817

APPENDICES

APPENDIX

For Previous SOTA results in Table I, we list the corresponding methods in Table III.
| Task | Dataset | full | all-way-0-shot | 10-way-0-shot | all-way-5-shot | 10-way-5-shot |
|------|---------|------|----------------|---------------|---------------|---------------|
| NER  | CoNLL03 | 94.60 [253] | 74.99 [162] | - | 83.25 [138] | - |
|      | OntoNotes5.0 | 92.30 [254] | - | - | 59.70 [159]\* | - |
|      | FewNERD | 70.90 [255] | - | - | 59.41 [159]\* | 79.00 [256] |
| RE   | NYT     | 93.50 [257] | - | - | - | - |
|      | TACREV  | 85.80 [258] | 59.40 [170]\* | - | 47.12 [180]\* | - |
|      | FewRel  | - | - | 84.20 [129] | - | 96.51 [259] |
| ED   | ACE05   | 83.65 [260] | 51.20 [138] | 54.50 [216] | 55.61 [134] | 64.80 [134] |
|      | MAVEN   | 79.09 [111] | 59.90 [216]\* | 36.86 [82]\* | 64.80 [16] | 93.06 [82]\† |
|      | FewEvent | 96.58 [260] | 58.14 [260] | 68.37 [82] | 60.67 [226]\† | 93.18 [82] |
| EAE  | ACE05   | 73.50 [134] | 31.20 [144]\* | 31.40 [134] | 45.90 [159]\* | 42.70 [134] |
|      | RAMS    | 59.66 [261] | - | - | 54.08 [159]\* | - |
|      | WikiEvents | 70.08 [261] | - | - | - | - |

TABLE III

The methods achieving the previous SOTA micro F1 performance, based on results in Table I. ⊿: LLM-based; ∗: type-specific prompting; ⋄: prompt-based meta learning; †: 45-way-5-shot; ‡: all-way-4-shot.