Large Language Models for Generative Information Extraction: A Survey

Anonymous ACL submission

Abstract

Information extraction (IE) aims to extract structural knowledge from plain natural language texts. Recently, generative Large Language Models (LLMs) have demonstrated remarkable capabilities in text understanding and generation. As a result, numerous works have been proposed to integrate LLMs for IE tasks based on a generative paradigm. To conduct a comprehensive systematic review and exploration of LLM efforts for IE tasks, in this study, we survey the most recent advancements in this field. We first present an extensive overview by categorizing these works in terms of various IE subtasks and learning paradigms, and then we empirically analyze the most advanced methods and discover the emerging trend of IE tasks with LLMs. Based on a thorough review conducted, we identify several insights in technique and promising research directions that deserve further exploration in future studies. We will maintain a public repository and consistently update related resources.

1 Introduction

Information Extraction (IE) is a crucial domain in natural language processing (NLP) that converts plain text into structured knowledge (e.g., entities, relations, and events), and serves as a foundational requirement for a wide range of downstream tasks, such as knowledge graph construction (Zhong et al., 2023), knowledge reasoning (Fu et al., 2019) and question answering (Srihari et al., 1999). Typical IE tasks consist of Named Entity Recognition (NER), Relation Extraction (RE) and Event Extraction (EE). Meanwhile, the emergence of large language models (LLMs) (e.g., GPT-4 (Achiam et al., 2023)) has greatly promoted the development of NLP, due to their extraordinary capabilities in text understanding and generation. Therefore, there has been a recent surge of interest in generative IE methods (Qi et al., 2023; Sainz et al., 2023) that adopt LLMs to generate structural information

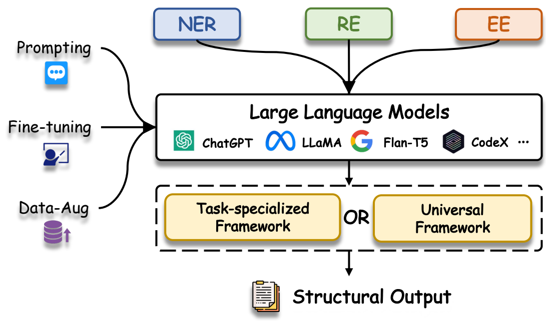

Figure 1: LLMs have been extensively explored for generative IE. These studies encompass various learning paradigms, specialized frameworks designed for a single subtask, and universal frameworks capable of addressing multiple subtasks simultaneously.

rather than extracting structural information from plain text. These methods have been proven to be more practical in real-world scenarios compared to discriminated methods (Chen et al., 2023a; Lou et al., 2023), as they efficiently handle schemas containing millions of entities without significant performance degradation (Josifoski et al., 2022).

On the one hand, LLMs have attracted significant attention from researchers in exploring their potentials for various scenarios of IE. In addition to excelling in individual IE tasks, LLMs possess a remarkable ability to effectively model various IE tasks in a universal format. This is conducted by capturing inter-task dependencies with instructive prompts, and achieves consistent performance (Lu et al., 2022; Sainz et al., 2023). On the other hand, recent works have shown the outstanding generalization of LLMs to not only learn from IE training data through fine-tuning (Paolini et al., 2021), but also extract information in few-shot and even zero-shot scenarios relying solely on in-context examples or instructions (Wei et al., 2023; Wang et al., 2023d). For above two groups of research works: 1) the universal frameworks for multiple tasks; 2) deficiency of training data scenarios, existing surveys (Nasar et al., 2021; Zhou et al., 2022a; Ye et al., 2022) do not fully explore them.

In this survey, we provide a comprehensive exploration of LLMs for generative IE, as illustrated in Figure 1. To achieve this, we categorize existing methods mainly using two taxonomies: (1) a taxonomy of numerous IE subtasks, which aims to classify different types of information that can be extracted individually or uniformly, and (2) a taxonomy of learning paradigms, which categorizes various novel approaches that utilize LLMs for generative IE. Furthermore, we present studies that focus on specific domains and analyze the performance of LLMs for IE. We also compare several representative methods to gain deeper understanding of their potentials and limitations, and provide insightful analysis on future directions. To the best of our knowledge, this is the first survey on generative IE with LLMs.

2 Preliminaries of Generative IE

In this section, we provide a formal definition of generative IE and summarize the IE subtasks. This generative IE survey primarily covers the tasks of NER, RE, and EE (Wang et al., 2023c; Sainz et al., 2023). The three types of IE tasks are formulated in a generative manner. Given an input text (e.g., sentence or document) with a sequence of n tokens $\mathcal{X} = [x_1, ..., x_n]$, a prompt $\mathcal{P}$, and the target extraction sequence $\mathcal{Y} = [y_1, ..., y_m]$, the objective is to maximize the conditional probability in an auto-regressive formulation:

$$p_\theta(\mathcal{Y}|\mathcal{X}, \mathcal{P}) = \prod_{i=1}^{m} p_\theta(y_i|\mathcal{X}, \mathcal{P}, y_{<i}), \qquad (1)$$

where θ donates the parameters of LLMs, which can be frozen or trainable. In the era of LLMs, several works have proposed appending extra prompts or instructions $\mathcal{P}$ to $\mathcal{X}$ to enhance the comprehensibility of the task for LLMs (Wang et al., 2023c). Even though the input text $\mathcal{X}$ remains the same, the target sequence varies for each task.

Named Entity Recognition (NER) includes two tasks: **Entity Identification** and **Entity Typing**. The former task is concerned with identifying spans of entities, and the latter task focuses on assigning types to these identified entities.

Relation Extraction (RE) may have different settings in different works. We categorize it using three terms following the literature (Lu et al., 2022; Wang et al., 2023c): (1) **Relation Classification** refers to classifying the relation type between two given entities; (2) **Relation Triplet** refers to identifying the relation type and the corresponding head and tail entity spans; (3) **Relation Strict** refers to giving the correct relation type, the span, and the type of head and tail entity.

Event Extraction (EE) can be divided into two subtasks (Wang et al., 2022a): (1) **Event Detection** (also known as Event Trigger Extraction in some works) aims to identify and classify the trigger word and type that most clearly represents the occurrence of an event. (2) **Event Arguments Extraction** aims to identify and classify arguments with specific roles in the events from the sentences.

3 LLMs for Different Information Extraction Tasks

In this section, we first present a introduction to the relevant LLM technologies for IE subtasks, including NER (§3.1), RE (§3.2), and EE (§3.3). We also conduct experimental analysis to evaluate the performance of various methods on representative datasets for three subtasks. Furthermore, we categorize universal IE frameworks into two categories: natural language (NL-LLMs) and code language (Code-LLMs), to discuss how they model the three distinct tasks using a unified paradigm (§3.4).

3.1 Named Entity Recognition

NER is a crucial component of IE and can be seen as a predecessor or subtask of RE and EE. It is also a fundamental task in other NLP tasks, thus attracting significant attention from researchers to explore new possibilities in the era of LLMs. Considering the gap between the sequence labeling and generation models, GPT-NER (Wang et al., 2023b) transformed NER into a generative task and proposed a self-verification strategy to rectify the mislabeling of NULL inputs as entities. Xie et al. (2023b) proposed a training-free self-improving framework that uses LLM to predict on the unlabeled corpus to obtain pseudo demonstrations, thereby enhancing the performance of LLM on zero-shot NER.

Table 1 shows the comparison of NER on five main datasets, which are obtained from their original papers. We can observe that: 1) the models in few-shot and zero-shot settings still have a huge performance gap behind the SFT and DA. 2) Even though there is little difference between backbones, the performance varies greatly between methods under the ICL paradigm. For example, GPT-NER opens up at least a 6% F1 value gap with other

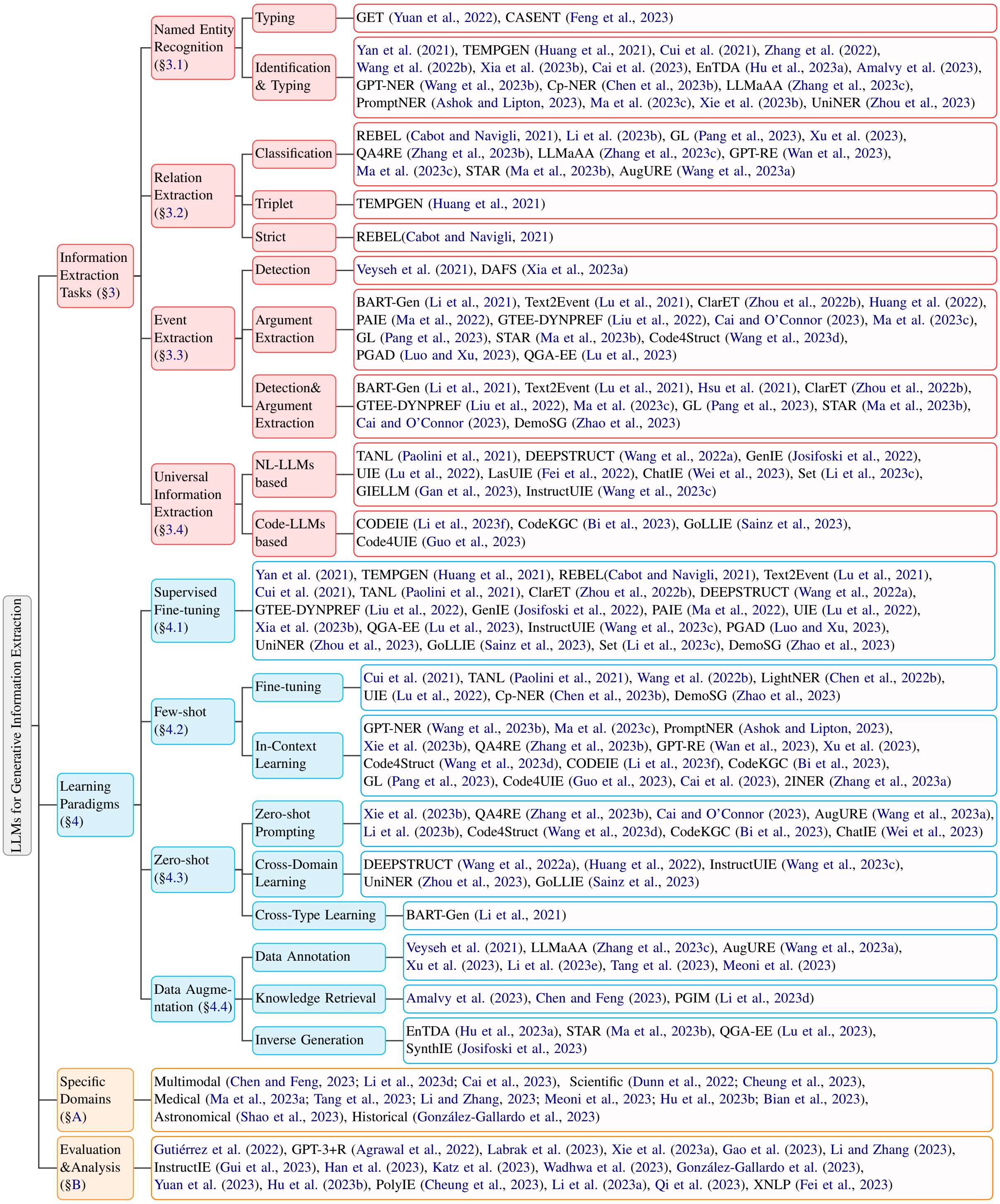

Figure 2: Taxonomy of research in generative IE using LLMs.

methods on each dataset, and up to about 19% higher. 3) Compared to ICL, there are only minor differences in performance between different models after SFT, even though the parameters in their backbones can differ by up to a few hundred times.

3.2 Relation Extraction

RE also plays an important role in IE, which usually has different setups in different studies as mentioned in Section 2. To address the poor performance of LLMs on RE tasks due to the low incidence of RE in instruction-tuning datasets, as indicated by Gutiérrez et al. (2022), QA4RE (Zhang et al., 2023b) introduced a framework that enhances LLMs' performance by aligning RE tasks with QA tasks. Due to the large number of predefined relation types and uncontrolled LLMs, Li et al. (2023e) proposed to integrate LLM with a natural language

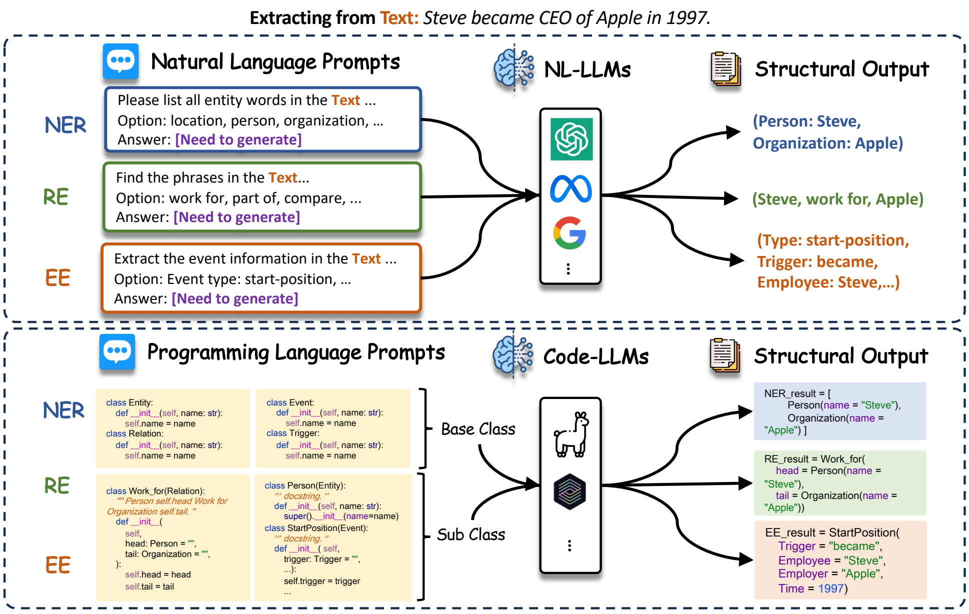

Figure 3: The comparison of prompts of NL-LLMs and Code-LLMs for universal IE. Both NL-based and code-based methods attempt to construct a universal schema, but they differ in terms of prompt format and the way they utilize the generation capabilities of LLMs. This figure is adopted from (Wang et al., 2023c) and (Guo et al., 2023).

inference module to generate relation triples, enhancing document-level relation datasets.

As shown in the Table 2 and 3, we statistically found that universal IE models are generally better solving harder Relation Strict problems due to learning the dependencies between multi-tasks (Paolini et al., 2021; Lu et al., 2022), while the task-specific methods solve simpler RE subtasks (e.g. relation classification). In addition, compared with NER, it can be found that the performance differences between models in RE are more obvious, indicating that the potential of LLM in RE task.

3.3 Event Extraction

Events can be defined as specific occurrences or incidents that happen in a given context. Recently, many studies (Liu et al., 2022; Lu et al., 2023) aim to understand events and capture their correlations by extracting event triggers and arguments using LLMs, which is essential for various reasoning tasks (Bhagavatula et al., 2020). For example, Code4Struct (Wang et al., 2023d) leveraged LLMs to translate text into code to tackle structured prediction tasks, using programming language features to introduce external knowledge and constraints through alignment between structure and code. PGAD (Luo and Xu, 2023) employed a text diffusion model to create a variety of context-aware prompt representations, enhancing both sentence-level and document-level event argument extraction by identifying multiple role-specific argument span queries and coordinating them with the context.

As can be seen from results of recent studies in Table 4, vast majority of current methods are based on SFT paradigm, and only a few methods that use LLMs for either zero-shot or few-shot learning. In addition, generative methods outperform discriminative ones by a wide margin, especially in metric for argument classification task, indicating the great potential of generative LLMs for EE.

3.4 Universal Information Extraction

Different IE tasks vary a lot, with different optimization objectives and task-specific schemas, requiring separate models to handle the complexity of different IE tasks, settings, and scenarios (Lu et al., 2022). As shown in Fig. 2, many works solely focus on a subtask of IE. However, recent advancements in LLMs have led to the proposal of a unified generative framework in several studies (Wang et al., 2023c; Sainz et al., 2023). This framework aims to model all IE tasks, capturing the common abilities of IE and learning the dependencies across multiple tasks. The prompt format for Uni-IE can typically be divided into natural language-based LLMs (NL-LLMs) and code-based LLMs (code-LLMs), as illustrated in Fig. 3.

NL-LLMs. NL-based methods unify all IE tasks in a universal natural language schema. For instance, UIE (Lu et al., 2022) proposed a unified text-to-structure generation framework that encodes extraction structures, and captured common IE abil-

Representative Model	Paradigm	Uni.	Backbone	ACE04	ACE05	CoNLL03	Onto. 5	GENIA
DEEPSTRUCT (Wang et al., 2022a)	CDL		GLM-10B	-	28.1	44.4	42.5	47.2
Xie et al. (2023b)	ZS Pr		GPT-3.5-turbo	-	32.27	74.51	-	52.06
CODEIE (Li et al., 2023f)	ICL	√	Code-davinci-002	55.29	54.82	82.32	-	-
Code4UIE (Guo et al., 2023)	ICL	√	Text-davinci-003	60.1	60.9	83.6	-	-
PromptNER (Ashok and Lipton, 2023)	ICL		GPT-4	-	-	83.48	-	58.44
Xie et al. (2023b)	ICL		GPT-3.5-turbo	-	55.54	84.51	-	58.72
GPT-NER (Wang et al., 2023b)	ICL		Text-davinci-003	74.2	73.59	90.91	82.2	64.42
TANL (Paolini et al., 2021)	SFT	√	T5-base	-	84.9	91.7	89.8	76.4
Cui et al. (2021)	SFT		BART	-	-	92.55	-	-
Yan et al. (2021)	SFT		BART-large	86.84	84.74	93.24	90.38	79.23
UIE (Lu et al., 2022)	SFT	√	T5-large	86.89	85.78	92.99	-	-
DEEPSTRUCT (Wang et al., 2022a)	SFT	√	GLM-10B	-	86.9	93.0	87.8	80.8
Xia et al. (2023b)	SFT		BART-large	87.63	86.22	93.48	90.63	79.49
InstructUIE (Gui et al., 2023)	SFT	√	Flan-T5-11B	-	86.66	92.94	90.19	74.71
UniNER (Zhou et al., 2023)	SFT		LLaMA-7B	87.5	87.6	-	89.1	80.6
GoLLIE (Sainz et al., 2023)	SFT	√	Code-LLaMA-34B	-	89.6	93.1	84.6	-
EnTDA (Hu et al., 2023a)	DA		T5-base	88.21	87.56	93.88	91.34	82.25
USM† (Lou et al., 2023)	SFT	√	RoBERTa-large	87.62	87.14	93.16	-	-
RexUIE† (Liu et al., 2023)	SFT	√	DeBERTa-v3-large	87.25	87.23	93.67	-	-
Mirror† (Zhu et al., 2023)	SFT	√	DeBERTa-v3-large	87.16	85.34	92.73	-	-

Table 1: Comparison of LLMs for named entity recognition (identification & typing) with the Micro-F1 metric (%). † indicates that the model is discriminative. We demonstrate some universal and discriminative models for comparison. Learning paradigms include Cross-Domain Learning (**CDL**), Zero-Shot Prompting (**ZS Pr**), In-Context Learning (**ICL**), Supervised Fine-Tuning (**SFT**), Data Augmentation (**DA**). **Uni.** denotes whether the model is universal. Onto. 5 denotes the OntoNotes 5.0. Details of datasets (§C) and backbones (§D) are presented in Appendix. The settings for all subsequent tables are consistent with this format.

ities through a structured extraction language. InstructUIE (Wang et al., 2023c) enhanced UIE by constructing expert-written instructions for fine-tuning LLMs to consistently model different IE tasks and capture the inter-task dependency. Additionally, ChatIE (Wei et al., 2023) explored the use of LLMs like ChatGPT (OpenAI, 2023) in zero-shot prompting, transforming the task into a multi-turn question-answering problem.

Code-LLMs. On the other hand, code-based methods unify IE tasks by generating code with a universal programming schema (Wang et al., 2023d). Code4UIE (Guo et al., 2023) proposed a universal retrieval-augmented code generation framework, which leverages Python classes to define schemas and uses in-context learning to generate codes that extract structural knowledge from texts. Besides, CodeKGC (Bi et al., 2023) leveraged the structural knowledge inherent in code and employed schema-aware prompts and rationale-enhanced generation to improve performance. To enable LLMs to adhere to guidelines out-of-the-box, GoLLIE (Sainz et al., 2023) enhanced zero-shot ability on unseen IE tasks by aligning with annotation guidelines.

In general, NL-LLMs are trained on a wide range of text and can understand and generate human language, which allows the prompts and instructions to be conciser and easier to design. However, NL-LLMs may produce unnatural outputs due to the distinct syntax and structure of IE tasks (Bi et al., 2023), which differ from the training data. Code, being a formalized language, possesses the inherent capability to accurately represent knowledge across diverse schemas, which makes it more suitable for structural prediction (Guo et al., 2023). But code-based methods often require a substantial amount of text to define a Python class (see Fig. 3), which in turn limits the sample size of the context. Through experimental comparison in Table 1, 2, and 4, we can observe that Uni-IE models after SFT outperform task-specific models in the NER, RE, and EE tasks for most datasets.

4 Learning Paradigms of LLMs for Generative IE

In this section, we categorize methods based on their learning paradigms, including **Supervised Fine-tuning** (§4.1, refers to further training LLMs on IE tasks using labeled data), **Few-shot Learning** (§4.2, refers to the generalization from a small number of labeled examples by training or in-context learning), **Zero-shot Learning** (§4.3, refers to generating answer without any training examples for the specific IE tasks), and **Data Augmentation** (§4.4, refers to enhancing information by applying various transformations to the existing data

Representative Model	Paradigm	Uni.	Backbone	NYT	ACE05	ADE	CoNLL04	SciERC
CodeKGC (Bi et al., 2023)	ZS Pr	√	Text-davinci-003	-	-	42.8	35.9	15.3
CODEIE (Li et al., 2023f)	ICL	√	Code-davinci-002	32.17	14.02	-	53.1	7.74
CodeKGC (Bi et al., 2023)	ICL	√	Text-davinci-003	-	-	64.6	49.8	24.0
Code4UIE (Guo et al., 2023)	ICL	√	Text-davinci-002	54.4	17.5	58.6	54.4	-
REBEL (Cabot and Navigli, 2021)	SFT		BART-large	91.96	-	82.21	75.35	-
UIE (Lu et al., 2022)	SFT	√	T5-large	-	66.06	-	75.0	36.53
InstructUIE (Wang et al., 2023c)	SFT	√	Flan-T5-11B	90.47	-	82.31	78.48	45.15
GoLLIE (Sainz et al., 2023)	SFT	√	Code-LLaMA-34B	-	70.1	-	-	-
USM† (Lou et al., 2023)	SFT	√	RoBERTa-large	-	67.88	-	78.84	37.36
RexUIE† (Liu et al., 2023)	SFT	√	DeBERTa-v3-large	-	64.87	-	78.39	38.37

Table 2: Comparison of LLMs for relation extraction with the "relation strict" (Lu et al., 2022) Micro-F1 metric (%). † indicates that the model is discriminative.

Representative Model	Paradigm	Uni.	Backbone	TACRED	Re-TACRED	TACREV	SemEval
QA4RE (Zhang et al., 2023b)	ZS Pr		Text-davinci-003	59.4	61.2	59.4	43.3
SUMASK (Li et al., 2023b)	ZS Pr		GPT-3.5-turbo-0301	79.6	73.8	75.1	-
GPT-RE (Wan et al., 2023)	ICL		Text-davinci-003	72.15	-	-	91.9
Xu et al. (2023)	ICL		Text-davinci-003	31.0	51.8	31.9	-
REBEL (Cabot and Navigli, 2021)	SFT		BART-large	-	90.36	-	-
Xu et al. (2023)	DA		Text-davinci-003	37.4	66.2	41.0	-

Table 3: Comparison of LLMs for relation classification with the Micro-F1 metric (%).

using LLMs), to highlight the commonly used approaches for adapting LLMs to IE.

4.1 Supervised Fine-tuning

Using all training data to fine-tune LLMs is the most common and promising method, which allows the model to capture the underlying structural patterns in the data, and generalize well to unseen samples. For example, DeepStruct (Wang et al., 2022a) introduced structure pre-training on a collection of task-agnostic corpora to enhance the structural understanding of language models. UniNER (Zhou et al., 2023) explored targeted distillation and mission-focused instruction tuning to train student models for broad applications, such as NER. GIELLM (Gan et al., 2023) fine-tuned LLMs using mixed datasets, which are collected to utilize the mutual reinforcement effect to enhance performance on multiple tasks.

4.2 Few-shot Learning

Few-shot learning has access to only a limited number of labeled examples, leading to challenges like overfitting and difficulty in capturing complex relationships (Huang et al., 2020). Fortunately, scaling up the parameters of LLMs gives them amazing generalization capabilities compared to small pre-trained models, allowing them to achieve excellent performance in few-shot settings (Li and Zhang, 2023; Ashok and Lipton, 2023). Paolini et al. (2021) proposed the Translation between Augmented Natural Languages (TANL) framework; Lu et al. (2022) proposed a text-to-structure generation framework (called UIE); and Chen et al. (2023b) proposed collaborative domain-prefix tuning for NER (called cp-NER). These methods have achieved state-of-the-art performance and demonstrated effectiveness in few-shot setting. Despite the success of LLMs, they face challenges in training-free IE because of the difference between sequence labeling and text-generation models (Gutiérrez et al., 2022). To overcome these limitations, GPT-NER (Wang et al., 2023b) introduced a self-verification strategy, while GPT-RE (Wan et al., 2023) enhanced task-aware representations and incorporates reasoning logic into enriched demonstrations. These approaches demonstrate how to effectively leverage the capabilities of GPT for in-context learning. CODEIE (Li et al., 2023f) and CodeKGC (Bi et al., 2023) showed that converting IE tasks into code generation tasks with code-style prompts and in-context examples leads to superior performance compared to NL-LLMs. This is because code-style prompts provide a more effective representation of structured output, enabling them to effectively handle the complex dependencies in natural language.

4.3 Zero-shot Learning

The main challenges in zero-shot learning lie in enabling the model to effectively generalize for tasks and domains that it has not been trained on, as well

Representative Model	Paradigm	Uni.	Backbone	Trg-I	Trg-C	Arg-I	Arg-C
Code4Struct (Wang et al., 2023d)	ZS Pr		Code-davinci-002	-	-	50.6	36.0
Code4UIE (Guo et al., 2023)	ICL	√	GPT-3.5-turbo-16k	-	37.4	-	21.3
Code4Struct (Wang et al., 2023d)	ICL		Code-davinci-002	-	-	62.1	58.5
TANL (Paolini et al., 2021)	SFT	√	T5-base	72.9	68.4	50.1	47.6
Text2Event (Lu et al., 2021)	SFT		T5-large	-	71.9	-	53.8
BART-Gen (Li et al., 2021)	SFT		BART-large	-	-	69.9	66.7
UIE (Lu et al., 2022)	SFT	√	T5-large	-	73.36	-	54.79
GTEE-DYNPREF (Liu et al., 2022)	SFT		BART-large	-	72.6	-	55.8
DEEPSTRUCT (Wang et al., 2022a)	SFT	√	GLM-10B	73.5	69.8	59.4	56.2
PAIE (Ma et al., 2022)	SFT		BART-large	-	-	75.7	72.7
PGAD (Luo and Xu, 2023)	SFT		BART-base	-	-	74.1	70.5
QGA-EE (Lu et al., 2023)	SFT		T5-large	-	-	75.0	72.8
InstructUIE (Wang et al., 2023c)	SFT	√	Flan-T5-11B	-	77.13	-	72.94
GoLLIE (Sainz et al., 2023)	SFT	√	Code-LLaMA-34B	-	71.9	-	68.6
USM† (Lou et al., 2023)	SFT	√	RoBERTa-large	-	72.41	-	55.83
RexUIE† (Liu et al., 2023)	SFT	√	DeBERTa-v3-large	-	75.17	-	59.15
Mirror† (Zhu et al., 2023)	SFT	√	DeBERTa-v3-large	-	74.44	-	55.88

Table 4: Comparison of Micro-F1 Values for Event Extraction on ACE05. Evaluation tasks include: Trigger Identification (Trg-I), Trigger Classification (Trg-C), Argument Identification (Arg-I), and Argument Classification (Arg-C). † indicates that the model is discriminative.

as aligning the pre-trained paradigm of LLMs. Due to the large amount of knowledge embedded within, LLMs show impressive abilities in zero-shot scenarios of unseen tasks (Kojima et al., 2022; Wei et al., 2023). To achieve zero-shot cross-domain generalization of LLMs in IE tasks, several works have been proposed (Sainz et al., 2023; Zhou et al., 2023; Wang et al., 2023c). These works offered a universal framework for modeling various IE tasks and domains, and introduced innovative training prompts, e.g., instruction (Wang et al., 2023c) and guidelines (Sainz et al., 2023), for learning and capturing the inter-task dependencies of known tasks and generalizing them to unseen tasks and domains. In terms of cross-type generalization, BART-Gen (Li et al., 2021) proposed a document-level neural model, by formulating EE task as conditional generation, resulting in better performance and excellent portability on unseen event types.

On the other hand, in order to improve the ability of LLMs under zero shot prompts (no need for further fine-tuning on IE tasks), QA4RE (Zhang et al., 2023b) and ChatIE (Wei et al., 2023) proposed to improve the performance of LLMs (like Flan-T5 (Chung et al., 2022) and GPT (Achiam et al., 2023)) on zero-shot IE tasks, by transforming IE into a multi-turn question-answering problem for aligning IE tasks with QA tasks. Li et al. (2023b) integrated the chain-of-thought approach and proposed the summarize-and-ask prompting to solve the challenge of ensuring the reliability of outputs from black box LLMs (Ma et al., 2023c).

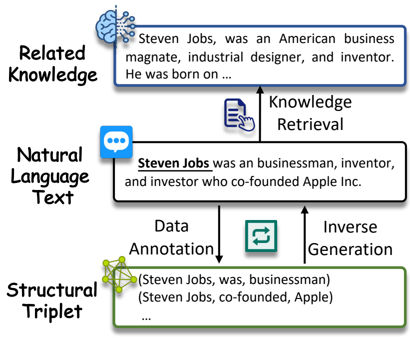

Figure 4: Comparison of data augmentation methods.

4.4 Data Augmentation

Data augmentation involves generating meaningful and diverse data to effectively enhance the training examples or information, while avoiding the introduction of unrealistic, misleading, and offset patterns. Recent powerful LLMs also demonstrate remarkable performance in data generation tasks (Whitehouse et al., 2023), which has attracted the attention of many researchers using LLMs to generate synthetic data for IE. It can be roughly divided into three strategies as shown in Fig. 4.

Data Annotation. This strategy directly generates labeled data using LLMs. For instance, Zhang et al. (2023c) proposed LLMaAA to improve accuracy and data efficiency by employing LLMs as annotators within an active learning loop, thereby optimizing both the annotation and training processes. AugURE (Wang et al., 2023a) employed within-sentence pairs augmentation and cross-sentence pairs extraction to enhance the diversity of positive pairs for unsupervised RE, and introduced margin loss for sentence pairs.

Knowledge Retrieval. This strategy retrieves relevant knowledge from LLMs for IE. PGIM (Li et al., 2023d) presented a two-stage framework for multimodal NER, which leverages ChatGPT as an implicit knowledge base to heuristically retrieve auxiliary knowledge for more efficient entity prediction. Amalvy et al. (2023) proposed to improve NER on long documents by generating a synthetic context retrieval training dataset, and training a neural context retriever.

Inverse Generation. This strategy prompts LLMs to produce natural text or questions based on structural data provided as input, aligning with training paradigm of LLMs. For example, SynthIE (Josifoski et al., 2023) showed that LLMs can create high-quality synthetic data for complex tasks by reversing the task direction, and train new models that outperformed previous benchmarks. Rather than relying on ground-truth targets, which limits generalizability and scalability, STAR (Ma et al., 2023b) generated structures from valid triggers and arguments, then generates passages with LLMs.

Overall, these strategies have their own advantages and disadvantages. While data annotation can directly meet task requirements, the ability of LLMs for structured generation still needs improvement. Knowledge retrieval can provide additional information about entities and relations, but it suffers from the hallucination problem and introduces noise. Inverse generation is aligned with the QA paradigm of LLMs. However, it requires structural data and there exists a gap between the generated pairs and the domain that needs to be addressed.

5 Future Directions

The development of LLMs for generative IE is still in its early stages, and there are numerous opportunities for improvement.

Universal IE. Previous generative IE methods and benchmarks are often tailored for specific domains or tasks, limiting their generalizability (Yuan et al., 2022). Although some unified methods (Lu et al., 2022) using LLMs have been proposed recently, they still suffer from certain limitations (e.g., long context input, and misalignment of structured output). Therefore, further development of universal IE frameworks that can adapt flexibly to different domains and tasks is a promising research direction (such as integrating the insights of task-specific models to assist in constructing universal models).

Low-Resource IE. The generative IE system with LLMs still encounters challenges in resource-limited scenarios (Li et al., 2023a). There is a need for further exploration of in-context learning of LLMs, particularly in terms of improving the selection of examples. Future research should prioritize the development of robust cross-domain learning techniques (Wang et al., 2023c), such as domain adaptation or multi-task learning, to leverage knowledge from resource-rich domains. Additionally, efficient data annotation strategies with LLMs should also be explored.

Prompt Design for IE. Designing effective instructions is considered to have a significant impact on the performance of LLMs (Qiao et al., 2022; Yin et al., 2023). One aspect of prompt design is to build input and output pairs that can better align with pre-training stage of LLMs (e.g., code generation) (Guo et al., 2023). Another aspect is optimizing the prompt for better model understanding and reasoning (e.g., Chain-of-Thought) (Li et al., 2023b), by encouraging LLMs to make logical inferences or explainable generation. Additionally, researchers can explore interactive prompt design (such as multi-turn QA) (Zhang et al., 2023b), where LLMs can iteratively refine or provide feedback on the generated extractions automatically.

Open IE. Open IE setting presents greater challenges for IE models, as it do not provide candidate label set and rely solely on the models' ability to comprehend the task. LLMs, with their knowledge and understanding abilities, have significant advantages in some Open IE tasks (Zhou et al., 2023). However, there are still instances of poor performance in more challenging tasks (Qi et al., 2023), which require further exploration by researchers.

6 Conclusion

In this survey, We first introduced the subtasks of IE and discussed some universal frameworks aiming to unify all IE tasks with LLMs. Additional theoretical and experimental analysis provided insightful exploration for these methods. Then we delved into different learning paradigms that apply LLMs for IE and demonstrate their potential for extracting information in specific domains. Finally, we analyzed the current challenges and presented potential future directions. We hope this survey can provide a valuable resource for researchers to explore more efficient utilization of LLMs for IE.

7 Limitations

This survey provides a comprehensive summary and analysis of the use of LLMs for generative IE, and points out related research directions. However, due to page and time constraints, there are still some limitations to this work.

Firstly, our work may have some omissions, such as some of the latest papers published on preprint websites. We will continue to update the latest related papers in our open-source repository. Furthermore, we summarized the prompt design of universal models based on code and natural language, but we have not summarized more prompt designs under few-shot and zero-shot scenarios. We actually discussed in the corresponding sections (e.g., Section 4.2 and 4.3) that some papers have proposed intricate and effective prompt designs. Notably, the Chain-of-thought method (Li et al., 2023b) and the multi-turn question answering (Zhang et al., 2023b) have demonstrated promising results in IE tasks. However, there is a limited number of such papers available, and their complexity does not support dedicating separate chapters to them. We have summarized the most common IE subtasks, but some similar technical directions are not included. For example, the settings of slot filling task (Dong et al., 2023; Li et al., 2023g) and NER tasks are similar, both of which are forms of sequence labeling. What is more, this survey mainly focuses on IE models with autoregressive LLMs as the backbone (e.g., LLaMA (Touvron et al., 2023) and ChatGPT (OpenAI, 2023)), thus lacks a summary and induction of discriminative IE models, (e.g., based on BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019)).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569.

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. Learning to rank context for named entity recognition using a synthetic dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382.

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations*.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2023. CodeKGC: Code language model for generative knowledge graph construction. *arXiv preprint arXiv:2304.09048*.

Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. 2023. Inspire the large language model by external knowledge on biomedical named entity recognition. *arXiv preprint arXiv:2309.12278*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. 33:1877–1901.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Chenran Cai, Qianlong Wang, Bin Liang, Bing Qin, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. In-context learning for few-shot multimodal named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2969–2979.

Erica Cai and Brendan O'Connor. 2023. A Monte Carlo language model pipeline for zero-shot sociopolitical event extraction. *arXiv preprint arXiv:2305.15051*.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. 2022a. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3329–3339.

Wei Chen, Lili Zhao, Pengfei Luo, Tong Xu, Yi Zheng, and Enhong Chen. 2023a. HEProto: A hierarchical enhancing protonet based on multi-task learning for few-shot named entity recognition. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 296–305.

Xiang Chen, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, Huajun Chen, and Ningyu Zhang. 2022b. LightNER: A lightweight tuning paradigm for low-resource NER via pluggable prompting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2374–2387.

Xiang Chen, Lei Li, Shuofei Qiao, Ningyu Zhang, Chuanqi Tan, Yong Jiang, Fei Huang, and Huajun Chen. 2023b. One model for all domains: Collaborative domain-prefix tuning for cross-domain NER. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5030–5038.

Jerry Junyang Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2023. PolyIE: A dataset of information extraction from polymer material scientific literature. *arXiv preprint arXiv:2311.07715*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *26th International Conference on Computational Linguistics*, pages 1169–1179.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

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

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3198–3213.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 837–840. European Language Resources Association.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Guanting Dong, Tingfeng Hui, Zhuoma GongQue, Jinxu Zhao, Daichi Guo, Gang Zhao, Keqing He, and Weiran Xu. 2023. Demonsf: A multi-task demonstration-based generative framework for noisy slot filling task. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10506–10518.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077.

Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022. LasUIE: Unifying information extraction with latent adaptive structure-aware generative language model. *Advances in Neural Information Processing Systems*, 35:15460–15475.

Hao Fei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. XNLP: An interactive demonstration system for universal structured nlp. *arXiv preprint arXiv:2308.01846*.

Yanlin Feng, Adithya Pratapa, and David R. Mortensen. 2023. Calibrated seq2seq models for efficient and generalizable ultra-fine entity typing. In *Findings of* the Association for Computational Linguistics, pages 15550–15560.

Cong Fu, Tong Chen, Meng Qu, Woojeong Jin, and Xiang Ren. 2019. Collaborative policy learning for open knowledge graph reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2672–2681.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2023. Giellm: Japanese general information extraction large language model utilizing mutual reinforcement effect. *arXiv preprint arXiv:2311.06838*.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of ChatGPT for event extraction. *arXiv preprint arXiv:2303.03836*.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, José G. Moreno, and Antoine Doucet. 2023. Yes but.. can chatgpt identify entities in historical documents? In *ACM/IEEE Joint Conference on Digital Libraries*, pages 184–189.

Runwei Guan, Ka Lok Man, Feifan Chen, Shanliang Yao, Rongsheng Hu, Xiaohui Zhu, Jeremy Smith, Eng Gee Lim, and Yutao Yue. 2023. FindVehicle and VehicleFinder: A NER dataset for natural language-based vehicle retrieval and a keyword-based cross-modal vehicle retrieval system. *arXiv preprint arXiv:2304.10893*.

Honghao Gui, Jintian Zhang, Hongbin Ye, and Ningyu Zhang. 2023. InstructIE: A Chinese instruction-based information extraction dataset. *arXiv preprint arXiv:2305.11527*.

Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, et al. 2023. Retrieval-augmented code generation for universal information extraction. *arXiv preprint arXiv:2311.02962*.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Informatics*, 45(5):885–892.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by ChatGPT? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2021. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908.

Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and S Yu Philip. 2023a. Entity-to-text based data augmentation for various named entity recognition tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9072–9087.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023b. Zero-shot clinical entity recognition using ChatGPT. *arXiv preprint arXiv:2303.16416*.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2020. Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978*.

Kuan-Hao Huang, I Hsu, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, et al. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4633–4646.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269.

Touvron Hugo, Martin Louis, Stone Kevin, and others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208.

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. GenIE: Generative information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *J. Biomed. Informatics*, 55:73–81.

Uri Katz, Matan Vetzler, Amir D. N. Cohen, and Yoav Goldberg. 2023. NERetrieve: Dataset for next generation named entity recognition and retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3340–3354.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, pages 180–182.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15.

Jin-Dong Kim, Yue Wang, and Yasunori Yamamoto. 2013. The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Aman Kumar and Binil Starly. 2022. "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *J. Intell. Manuf.*, 33(8):2393–2407.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2023. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv preprint arXiv:2307.12114*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023b. Revisiting large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892.

Jiangnan Li, Yice Zhang, Bin Liang, Kam-Fai Wong, and Ruifeng Xu. 2023c. Set learning for generative information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13043–13052.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023d. Prompting ChatGPT in MNER: enhanced multimodal named entity recognition with auxiliary refined knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2787–2802.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023e. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. *arXiv preprint arXiv:2311.07314*.

Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023f. CodeIE: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908.

Xuefeng Li, Liwen Wang, Guanting Dong, Keqing He, Jinzheng Zhao, Hao Lei, Jiachi Liu, and Weiran Xu. 2023g. Generative zero-shot prompt learning for cross-domain slot filling with inverse prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 825–834.

Chengyuan Liu, Fubang Zhao, Yangyang Kang, Jingyuan Zhang, Xiang Zhou, Changlong Sun, Kun Kuang, and Fei Wu. 2023. RexUIE: A recursive method with explicit schema instructor for universal information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15342–15359.

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

Xiao Liu, He-Yan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 13318–13326.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. Event extraction as question generation and answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1666–1688.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. 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)*, pages 2795–2806.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Lei Luo and Yajing Xu. 2023. Context-aware prompt for generation-based event argument extraction with diffusion models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1717–1725.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023a. DICE: data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917. Association for Computational Linguistics.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2023b. STAR: Boosting low-resource event extraction by structure-to-text data generation with large language models. *arXiv preprint arXiv:2305.15090*.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023c. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190.

Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee M. Christensen, David Martínez, Liadh Kelly, Lorraine Goeuriot, Noémie Elhadad, Sameer Pradhan, Guergana K. Savova, and Wendy W. Chapman. 2014. Task 2: ShARe/CLEF eHealth evaluation lab 2014. In *Working Notes for CLEF 2014 Conference*, volume 1180, pages 31–42.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys*, 54(1):1–39.

OpenAI. 2023. Introduce ChatGPT. *OpenAI blog*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15372–15389.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations*.

Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova. 2013a. Task 1: ShARe/CLEF eHealth evaluation lab 2013. In *Working Notes for CLEF 2013 Conference*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013b. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013*, pages 143–152.

Ji Qi, Chuchun Zhang, Xiaozhi Wang, Kaisheng Zeng, Jifan Yu, Jinxin Liu, Lei Hou, Juanzi Li, and Xu Bin. 2023. Preserving knowledge invariance: Rethinking robustness evaluation of open information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 5876–5890.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323, pages 148–163.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning*, pages 1–8.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code LLama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. GoLLIE: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning*, pages 142–147.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference*, pages 8749–8757.

Wujun Shao, Yaohua Hu, Pengli Ji, Xiaoran Yan, Dongwei Fan, and Rui Zhang. 2023. Prompt-NER: Zero-shot named entity recognition in astronomy literature via large language models. *arXiv preprint arXiv:2310.17892*.

Rohini Srihari, Wei Li, and X Li. 1999. Information extraction supported question answering. In *TREC*.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 13843–13850. AAAI Press.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron C. Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and et al. Azhar, Faisa. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Asahi Ushio, Francesco Barbieri, Vítor Silva, Leonardo Neves, and José Camacho-Collados. 2022. Named entity recognition in twitter: A dataset and analysis on short-term temporal shifts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers*, pages 309–319.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash GPT-2 power for event detection. 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)*, pages 6271–6282.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *https://catalog.ldc.upenn.edu/LDC2006T06*.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: in-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-parallel implementation of Transformer language model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022a. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022b. InstructionNER: A multi-task instruction-based generative framework for few-shot NER. *arXiv preprint arXiv:2203.03903*.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023a. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

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

Xingyao Wang, Sha Li, and Heng Ji. 2023d. Code4Struct: Code generation for few-shot event structure prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3640–3663.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022c. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5153–5162.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*.

Nan Xia, Hang Yu, Yin Wang, Junyu Xuan, and Xiangfeng Luo. 2023a. DAFS: a domain aware few shot generative model for event detection. *Machine Learning*, 112(3):1011–1031.

Yu Xia, Yongwei Zhao, Wenhao Wu, and Sujian Li. 2023b. Debiasing generative named entity recognition by calibrating sequence likelihood. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1137–1148.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023a. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023b. Self-improving for zero-shot named entity recognition with large language models. *arXiv preprint arXiv:2311.08921*.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing*, pages 190–200. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 1: Long Papers)*, pages 5808–5822.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 764–777.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. *arXiv preprint arXiv:2210.12714*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang, and Yanghua Xiao. 2022. Generative entity typing with curriculum learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3061–3073.

Jiasheng Zhang, Xikai Liu, Xinyi Lai, Yan Gao, Shusen Wang, Yao Hu, and Yiqing Lin. 2023a. 2iner: Instructive and in-context learning on few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3940–3951.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023b. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5674–5681.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023c. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103.

Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. 2022. De-bias for generative extraction in unified NER task. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 808–818.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Gang Zhao, Xiaocheng Gong, Xinjie Yang, Guanting Dong, Shudong Lu, and Si Li. 2023. DemoSG: Demonstration-enhanced schema-guided generation for low-resource event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1805–1816.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *arXiv preprint arXiv:2302.05019*.

Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, Haiyang Yu, Jian Sun, and Yongbin Li. 2022a. A survey on neural open information extraction: Current status and future directions. *arXiv preprint arXiv:2205.11725*.

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

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022b. Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2559–2575.

Tong Zhu, Junfei Ren, Zijian Yu, Mengsong Wu, Guoliang Zhang, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2023. Mirror: A universal framework for various information extraction tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8861–8876.

A Specific Domains

It is non-ignorable that LLMs have tremendous potential for extracting information from some specific domains, such as mulitmodal (Chen and Feng, 2023; Li et al., 2023d), medical (Tang et al., 2023; Ma et al., 2023a) and scientific (Dunn et al., 2022; Cheung et al., 2023) information.

Multimodal. Chen and Feng (2023) introduced a conditional prompt distillation method that enhances a model's reasoning ability by combining text-image pairs with chain-of-thought knowledge from LLMs, significantly improving performance in multimodal NER and multimodal RE.

Medical. Tang et al. (2023) explored the potential of LLMs in the field of clinical text mining and proposed a novel training approach, which leverages synthetic data to enhance performance and address privacy issues.

Scientific. Dunn et al. (2022) presented a sequence-to-sequence approach by using GPT-3 for joint NER and RE from complex scientific text, demonstrating its effectiveness in extracting complex scientific knowledge in material chemistry.

B Evaluation & Analysis

Despite the great success of LLMs in various natural language processing tasks, their performance in the field of information extraction still have room for improvement (Han et al., 2023). To alleviate this problem, recent research has explored the capabilities of LLMs with respect to the major subtasks of IE, i.e., NER (Xie et al., 2023a; Li and Zhang, 2023), RE (Wadhwa et al., 2023; Yuan et al., 2023), and EE (Gao et al., 2023). Considering the superior reasoning capabilities of LLMs, Xie et al. (2023a) proposed four reasoning strategies for NER, which are designed to simulate ChatGPT's potential on zero-shot NER. Wadhwa et al. (2023) explored the use of LLMs for RE and found that few-shot prompting with GPT-3 achieves near SOTA performance, while Flan-T5 can be improved with chain-of-thought style explanations generated via GPT-3. For EE task, Gao et al. (2023) showed that ChatGPT still struggles with it due to the need for complex instructions and a lack of robustness.

Along this line, some researchers performed a more comprehensive analysis of LLMs by evaluating multiple IE subtasks simultaneously. Li et al. (2023a) evaluated ChatGPT's overall ability on IE, including performance, explainability, calibration, and faithfulness. They found that ChatGPT mostly performs worse than BERT-based models in the standard IE setting, but excellently in the OpenIE setting. Furthermore, Han et al. (2023) introduced a soft-matching strategy for a more precise evaluation and identified "unannotated spans" as the predominant error type, highlighting potential issues with data annotation quality.

C Benchmarks

As shown in Table 5, we compiled a comprehensive collection of benchmarks covering various domains and tasks, to provide researchers with a valuable resource that they can query and reference as needed. Moreover, we also summarized the download links for each dataset in our open source repository.

D Backbones

We briefly describe some backbones that are commonly used in the field of generative information extraction, which is shown in Table 6.

Table 5: Statistics of common datasets for information extraction. ∗ denotes the dataset is multimodal. # refers to the number of categories or sentences. The data in the table is partially referenced from InstructUIE (Gui et al., 2023).

Task	Dataset	Domain	#Class	#Train	#Val	#Test
NER	ACE04 (Doddington et al., 2004)	News	7	6,202	745	812
	ACE05 (Walker et al., 2006)	News	7	7,299	971	1,060
	BC5CDR (Li et al., 2016)	Biomedical	2	4,560	4,581	4,797
	Broad Twitter Corpus (Derczynski et al., 2016)	Social Media	3	6,338	1,001	2,000
	CADEC (Karimi et al., 2015)	Biomedical	1	5,340	1,097	1,160
	CoNLL03 (Sang and De Meulder, 2003)	News	4	14,041	3,250	3,453
	CoNLLpp (Wang et al., 2019)	News	4	14,041	3,250	3,453
	CrossNER-AI (Liu et al., 2021)	Artificial Intelligence	14	100	350	431
	CrossNER-Literature (Liu et al., 2021)	Literary	12	100	400	416
	CrossNER-Music (Liu et al., 2021)	Musical	13	100	380	465
	CrossNER-Politics (Liu et al., 2021)	Political	9	199	540	650
	CrossNER-Science (Liu et al., 2021)	Scientific	17	200	450	543
	FabNER (Kumar and Starly, 2022)	Scientific	12	9,435	2,182	2,064
	Few-NERD (Ding et al., 2021)	General	66	131,767	18,824	37,468
	FindVehicle (Guan et al., 2023)	Traffic	21	21,565	20,777	20,777
	GENIA (Kim et al., 2003)	Biomedical	5	15,023	1,669	1,854
	HarveyNER (Chen et al., 2022a)	Social Media	4	3,967	1,301	1,303
	MIT-Movie (Liu et al., 2013)	Social Media	12	9,774	2,442	2,442
	MIT-Restaurant (Liu et al., 2013)	Social Media	8	7,659	1,520	1,520
	MultiNERD (Tedeschi and Navigli, 2022)	Wikipedia	16	134,144	10,000	10,000
	NCBI (Doğan et al., 2014)	Biomedical	4	5,432	923	940
	OntoNotes 5.0 (Pradhan et al., 2013b)	General	18	59,924	8,528	8,262
	ShARe13 (Pradhan et al., 2013a)	Biomedical	1	8,508	12,050	9,009
	ShARe14 (Mowery et al., 2014)	Biomedical	1	17,404	1,360	15,850
	SNAP* (Lu et al., 2018)	Social Media	4	4,290	1,432	1,459
	TTC (Rijhwani and Preotiuc-Pietro, 2020)	Social Meida	3	10,000	500	1,500
	Tweebank-NER (Jiang et al., 2022)	Social Media	4	1,639	710	1,201
	Twitter2015* (Zhang et al., 2018)	Social Media	4	4,000	1,000	3,357
	Twitter2017* (Lu et al., 2018)	Social Media	4	3,373	723	723
	TwitterNER7 (Ushio et al., 2022)	Social Media	7	7,111	886	576
	WikiDiverse* (Wang et al., 2022c)	News	13	6,312	755	757
	WNUT2017 (Derczynski et al., 2017)	Social Media	6	3,394	1,009	1,287
RE	ACE05 (Walker et al., 2006)	News	7	10,051	2,420	2,050
	ADE (Gurulingappa et al., 2012)	Biomedical	1	3,417	427	428
	CoNLL04 (Roth and Yih, 2004)	News	5	922	231	288
	DocRED (Yao et al., 2019)	Wikipedia	96	3,008	300	700
	MNRE* (Zheng et al., 2021)	Social Media	23	12,247	1,624	1,614
	NYT (Riedel et al., 2010)	News	24	56,196	5,000	5,000
	Re-TACRED (Stoica et al., 2021)	News	40	58,465	19,584	13,418
	SciERC (Luan et al., 2018)	Scientific	7	1,366	187	397
	SemEval2010 (Hendrickx et al., 2010)	General	19	6,507	1,493	2,717
	TACRED (Zhang et al., 2017)	News	42	68,124	22,631	15,509
	TACREV (Alt et al., 2020)	News	42	68,124	22,631	15,509
EE	ACE05 (Walker et al., 2006)	News	33/22	17,172	923	832
	CASIE (Satyapanich et al., 2020)	Cybersecurity	5/26	11,189	1,778	3,208
	GENIA11 (Kim et al., 2011)	Biomedical	9/11	8,730	1,091	1,092
	GENIA13 (Kim et al., 2013)	Biomedical	13/7	4,000	500	500
	PHEE (Sun et al., 2022)	Biomedical	2/16	2,898	961	968
	RAMS (Ebner et al., 2020)	News	139/65	7,329	924	871
	bart-gen (Li et al., 2021)	Wikipedia	50/59	5,262	378	492

Table 6: The common backbones for generative information extraction. We mark the commonly used base and large versions for better reference.

Series	Model	Size	Base Model	Open Source	Instruction Tuning	RLHF
BART	BART	140M (base), 400M (large)	-	√	-	-
T5	T5 (Raffel et al., 2020)	60M, 220M (base), 770M (large), 3B, 11B	-	√	-	-
	mT5 (Xue et al., 2021)	300M, 580M (base), 1.2B (large), 3.7B, 13B	-	√	-	-
	Flan-T5 (Chung et al., 2022)	80M, 250M (base), 780M (large), 3B, 11B	T5	√	√	-
GLM	GLM (Du et al., 2022)	110M (base), 335M (large), 410M, 515M, 2B, 10B	-	√	-	-
LLaMA	LLaMA (Touvron et al., 2023)	7B, 13B, 33B, 65B	-	√	-	-
	Alpaca (Taori et al., 2023)	7B, 13B	LLaMA	√	√	-
	Vicuna (Chiang et al., 2023)	7B, 13B	LLaMA	√	√	-
	LLaMA2 (Hugo et al., 2023)	7B, 13B, 70B	-	√	-	-
	LLaMA2-chat (Hugo et al., 2023)	7B, 13B, 70B	LLaMA2	√	√	√
	Code-LLaMA (Roziere et al., 2023)	7B, 13B, 34B	LLaMA2	√	-	-
GPT	GPT-2 (Radford et al., 2019)	117M, 345M, 762M, 1.5B	-	√	-	-
	GPT-3 (Brown et al., 2020)	175B	-	-	-	-
	GPT-J (Wang, 2021)	6B	GPT-3	√	-	-
	Code-davinci-002 (Ouyang et al., 2022)	-	GPT-3	-	√	-
	Text-davinci-002 (Ouyang et al., 2022)	-	GPT-3	-	√	-
	Text-davinci-003 (Ouyang et al., 2022)	-	GPT-3	-	√	√
	GPT-3.5-turbo series (OpenAI, 2023)	-	-	-	√	√
	GPT-4 series (Achiam et al., 2023)	-	-	-	√	√