General Collaborative Framework between Large Language Model and Experts for Universal Information Extraction

Anonymous ACL submission

Abstract

Recently, unified information extraction have been widely concerned NLP community, which aims at using a unified paradigm to perform various information extraction tasks. However, they inevitably suffering from some thorny problems such as noise interference, abstract label semantics, and diverse span granularity. In this paper, First of all, we start by presenting three problematic assumptions that exist in previous research works from a unified information extraction perspective. These problems severely hinder the development of information extraction models. Furthermore, to solve these problems, we propose the **G**eneral **C**ollaborative **I**nformation **E**xtraction framework for universal information extraction. Specifically, GCIE consists of a general Recognizer for identifying predefined types and multiple task-specific Experts for extracting spans. The Recognizer is a large language model, while the Expert is a series of smaller language models, and they collaborate in a pipeline to achieve unified or task-specific information extraction. Empirical experiments on 6 IE tasks and 13 datasets, under supervised and few-shot settings, validate the effectiveness and generality of our approach.

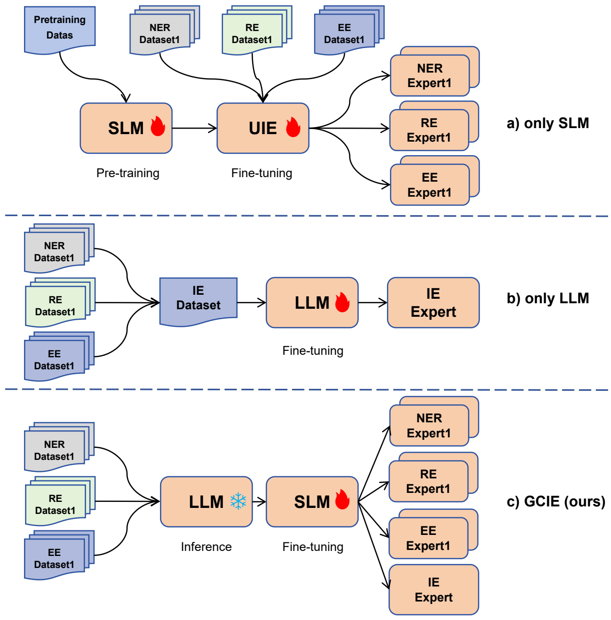

Figure 1: The difference of GCIE with other main paradigms for unified information extraction. a) pre-training and fine-tuing with SLM; b) instruction fine-tuing with LLM; c) inference with LLM and fine-tuning with SLM.

1 Introduction

Information Extraction (IE) aims to extract structured information from unstructured text (Andersen et al., 1992; Grishman, 2019). This is a complex task consisting of a series of subtasks, such as named entity recognition, relation extraction, entity linking, aspect-based sentiment analysis, event extraction, etc (Muslea, 1999). Due to its various targets (entity, relation, event, etc.), heterogeneous structures (spans, triplets, records, etc.), traditional methods for IE usually develop task-specialized architectures and processes, which is commonly required elaborate manual design (Grishman and Sundheim, 1996; Ji and Grishman, 2011). These task-specialized solutions greatly hinder the rapid architecture development, So another line of research of IE focus on resolving multiple subtasks using a universal model, such as recent works (Peng et al., 2023; Ping et al., 2023).

(Lu et al., 2022) proposed the unified structured generation model (UIE) for 4 IE subtasks, which generates structured texts according to task targets on the foundation of T5 (Raffel et al., 2020), offering a new thread and paradigm to IE research. However, UIE still has three significant issues that remain unresolved. For example, the noise interference introduced by negative samples during model training. Unlike typical NLP tasks, the data used for information extraction tasks exhibit imbalanced label quantities across various types, with a much larger number of negative samples compared to positive ones (Huang et al., 2020; Dong et al., 2021;

Liu et al., 2023). On the another line, (Lin et al., 2020; Lou et al., 2023; Ping et al., 2023) use extractive models to accomplish universal information extraction via heterogeneous decoding process on different subtasks. Just like UIE, one major challenge with them is getting validate token representation, especially of label prompt. However, unlike large language model such as GPT-3, PaLM, LLaMA, etc (Patel et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023), smaller language model (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020) have not sufficient power of making sense of label when type phrase is very abstract. For example, "Attack" is a event type in ACE05-Evt, representing a series of conflict events such as wars, coups, strikes, terrorist attacks, etc, not just its literal meaning.

Witnessing the outstanding performance of massively large language models in traditional tasks of NLP, several LLM-based method for information extraction had been proposed (Zhou et al., 2023; Wang et al., 2023b, 2022a; Wadhwa et al., 2023a; Gui et al., 2023; Wang et al., 2023c). But there is still no consensus on the trade-off between effectiveness and efficiency, due to considerable gap in IE tasks (Han et al., 2023) and overhead of training LLM.

In this paper, we dedicated to analysing these key problems and trying to find a solution. Based on the above statement and our exploration, 3 main factors that affect the performance of information extraction models are summarized as : 1) Noisy imbalanced data: a large of negative samples and long-tail data distribution. 2) Abstract label type: obscure type phrases to understanding by LMs. 3) Diverse span granularity: different span identification criteria in data annotations. Accordingly, We assume that anti-interference, label-understanding, and span-identification are the primary capabilities of information extraction models, corresponding to the 3 vital problems mentioned above. After that, We have proposed a general collaborative framework with these capabilities for universal information extraction, composed of a Recognizer and multiple Experts. Specifically, Recognizer is a LLM good at anti-interference and label-understanding, recognizing label types and filtering negative samples. And Expert receives type indicator as prompt to generate structured texts, which are fine-tuned on noise-free data for specific IE tasks. Recognizer and Experts work together in a pipeline to generate general schema for universal IE tasks, as shown in Figure 1. Different from previous research, our approach focuses more on solving the aforementioned problems and achieving further performance improvement by simultaneously utilizing the advantages of LLM and SLM.

To verify the effectiveness and generality of GCIE, extensive experiments including 6 IE subtasks with 13 datasets are conducted on supervised and few-shot settings. Experimental results show that, GCIE achieves state-of-the-art performance on several datasets on the supervised settings, and significantly outperforms all baselines on the few-shot settings. All indicate the combination of SLM and LLM result in stronger information extraction capabilities.

In summary, Our main contributions are summarized as follows:

1) We summarize three primary abilities of IE models corresponding to three vital problems in IE tasks and reveal the reasons that hinder the performance improvement of information extraction models.

2) We propose a general collaborative framework for information extraction, which leverages the complementary advantages of LLM and SLM to further improve the performance of unified structured generation information extraction models.

3) We design prompts for accurate type recognition of LLM and self-correction learning strategies for effective Expert training.

4) We conduct extensive experiments on 6 IE tasks and 13 datasets, exploring the overall performance of GCIE under supervised and few-shot settings, as well as the main factors influencing the performance of Recognizer and Expert. These experiments confirm the effectiveness and generality of GCIE.

2 Key Capabilities for Information Extraction

In this section, We summarize the necessary conditions for addressing the challenges in IE tasks as three key capabilities and explain why an excellent IE model should possess both of these capabilities.

Anti-interference describes the robustness of an IE model to data distribution with noise. In real world, there are little information or lacking of annotation in many texts that are usually called negative samples. For example, both ACE2005 and SciERC have quite a few negative samples,

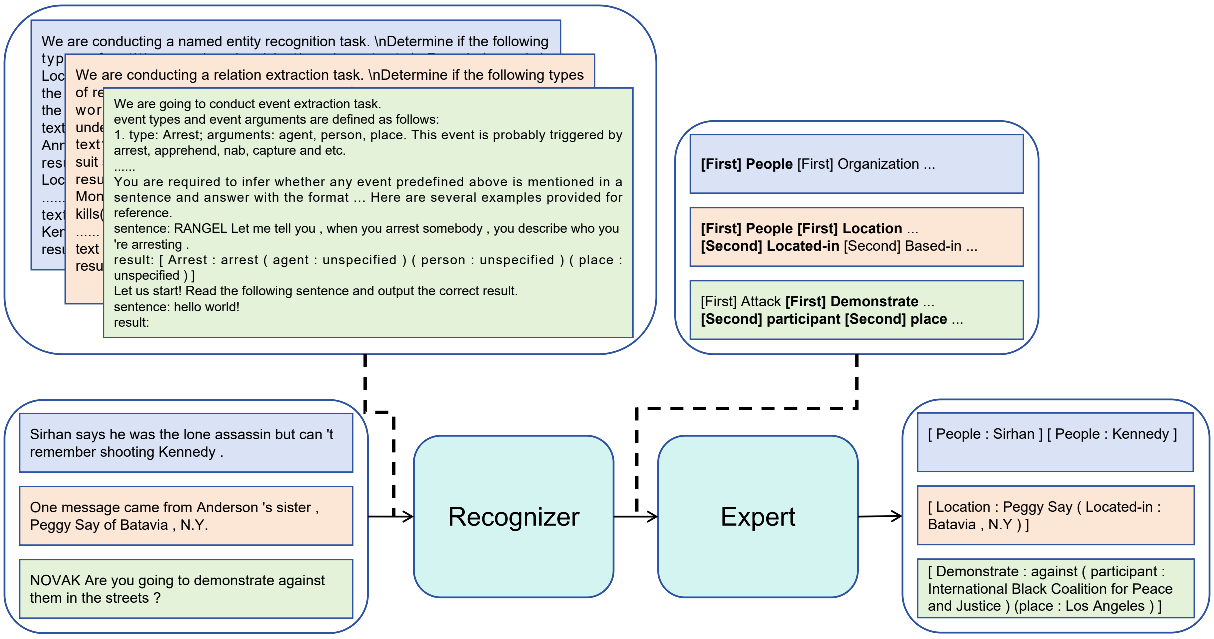

Figure 2: The overall framework of GCIE. In the prompts of Expert, types recognized by Recognizer are marked in bold.

involving event extraction and relation extraction tasks. To verify the harmful impact of negative samples made on IE models, we conduct several experiments (See Appendix B) on negative samples identification with three different paradigms including context-based large language model and classic generative IE models. In our observation, all fine-tuning-based IE models struggle to distinguish negative samples with those are similar in content but informative samples, which is attributed to overfitting. In contrast, few-shot contexts enable LLM to identify negative samples excellently. Therefore, we consider Anti-interference as a necessary condition for developing better IE models.

Label-understanding describes the semantic understanding ability of model to predefined label. In recent years, many research works unlocked label semantic understanding abilities of PLM via prompt learning across a series of NLP tasks, such as summary, text classification, text generation, sentiment analysis and few-shot NER (Narayan et al., 2021; Zou et al., 2021; Seoh et al., 2021; Schick and Schütze, 2021; Ma et al., 2022). However, these phenomena only emerge in NLP tasks with simple label words like positive, great, person, etc. More abstractive and polysemous label words are too obscure to understand for common language models. In our ablation experiments, We observed that model's performances vary to different degrees depending on different types when we take place of some type words with capitals or other lexical items. At least to some extent, SLM does more about statistical mapping between labels and the text than understanding and generating through abstract label semantics in comparison to LLM.

Span-identification describes the capacity of completely identifying phases that probably are entities, event triggers or event arguments. To investigate this ability of IE models, we test T5-base, T5-large and Claude2 on tasks of entity and trigger mentions extraction. Just as (Han et al., 2023) report, this ability of LLM is far less than that of fine-tuned SLM. The reason for this phenomenon is obviously that different dataset with different annotation styles have various span granularity, such as "man" - "the man", "hospital in Boston" - "hospital", "2 soldiers" - "soldiers" and etc. Under strict evaluation metrics, LLM without any training process has difficult in competing with fine-tuned SLM.

3 General Collaborative Framework

Based on above statement, we introduce a two-stage (Recognition and Generation) general collaborative framework combining LLM and SLM to possess abilities of Anti-interference, Label-understanding and Span-identification for universal information extraction tasks.

3.1 Schema Definition

Inspired by previous researches, we format all IE subtasks as unified structure generation (see appendix E). Formally, given a sentence s as input, our GCIE outputs structure schema o, which consists of tokens coming from label collection, in-context collection and structure collection. Figure 3 demonstrates several examples for this unified schema. Wherein the label collection includes predefined label type words, and the in-context collection is made up of input tokens. Different from previous studies, We use two symbols to hold the primary and secondary structures respectively. The output format is used in both the two stages of Recognition and Generation.

3.2 Framework Architecture

Our framework consists of Recognizer (a black-box LLM only used for inference) and Expert (a fine-tuned SLM), shown as Figure 2. In detail, Recognizer receives a sentence s and a task-special instruction that contains examples e and question q as the input. With a few examples of input-output pairs to refer, Recognizer answers the question in the same format. The result given by Recognizer can be write as follow:

$$a = Recognizer(s, q|e) \tag{1}$$

where $a = \{(typ_1, val_1), ..., (typ_n, val_n)\}$ is a dict with n type words and binary values as items, indicating whether a predefined type exists in the sentence. q is a task-special question to query LLM for a rational answer and $e = \{(text_1, result_1), ..., (text_k, result_k)\}$ is k-shot examples.

In the question designed by us, all predefined label types were represented by a single word or a phrase with a short description as interpretation. Through linking these interpretation with examples, instead of hard tokens used in SLM-only methods, LLM understands the actual semantics of each type vocabulary better and never minds overfitting problems caused by model training. It is important that we select examples from train set as diversely and comprehensively as possible. In this way, LLM can recognize available types in the sentence stably and precisely.

After recognition of LLM, phrases of confident types as the prompts are concatenated with text tokens as inputs of Expert, denoted respectively as $p = \{p_1, p_2, ..., p_{|p|}\}$ and $t = \{t_1, t_2, ..., t_{|t|}\}$. p_i comes from $a_{val_i=1}$. The form of p for each dataset is listed in Appendix. Theoretically, any auto-regression and encoder-decoder language model could be used as Expert, which predict conditional probability $p(y_i|y_{<i}, p, t)$ of the next token y_i, given the in-context and input. Finally, when Expert finishes prediction when it generate the end symbol, appropriate sampling technique is applied to get the final output sequence o, totally write as follow:

$$o = Expert(p, t) \tag{2}$$

$$o_i = Decoder(o_{<i}, p, t) \tag{3}$$

where $o = o_1, o_2, ..., o_{|o|}$ is the result of sampling with structured schema we defined above and $|o|$ is the length of output sequence. $Decoder(\cdot)$ is the decoder of Expert.

Due to structured schema rather than natural language text, previous study applies constrained-decode for the process of schema generation to accomplish controllable structure generation (Lu et al., 2021). This is an optional solution for ours but no significant effects are observed compared with greedy search and beam search. One trick we consider very important, is the uniqueness of type phrases. For instance, We suggest type word "method" is substituted by "Methods", because the "method" in text typically is a entity with type of "Generic".

3.3 Expert Learning

To be equipped with Span-identification ability, Expert requires a fine-tuning process. At present, we consider multiple feasible training plan which produces two bifurcation points. One is whether supervised datas from gold label or from Recognizer prediction are used to training. Another is whether multiple task-specific Experts or a unified Expert for all IE tasks are maintain. We conducted a thorough investigation of these issues in our experiments. For simplicity, we assume $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ uniformly represent arbitrary training dataset. Therefore a mostly straightforward way to optimize parameters is minimizing the negative logarithmic likelihood expectation on train set:

$$\mathcal{L} = \sum_{(x,y)\in\mathcal{D}} -logp(y|x, p; \theta) \tag{4}$$

where p is type tokens from Recognizer prediction or gold label and θ is trainable parameters in

Model	NER			RETriplet	NER&RE					
	CoNLL03	GENIA	ACE05-Ent	NYT	CoNLL04		SciERC		ACE05-Rel	
	Ent	Ent	Ent	Ent	Ent	Rel	Ent	Rel	Ent	Rel
(Shen et al., 2022)	92.87	**81.77**	87.42	-	-	-	-	-	-	-
(Li et al., 2022)	93.07	81.39	86.79	-	-	-	-	-	-	-
(Yan et al., 2021)	-	-	-	92.40	-	-	66.80	38.40	89.00	66.80
(Tang et al., 2022)	-	-	-	93.70	-	-	-	-	-	-
(Shen et al., 2021)	-	-	-	-	90.30	72.35	-	-	87.61	62.77
(Lu et al., 2022)†	92.99	-	85.78	-	-	75.00	-	36.53	-	66.06
(Lou et al., 2023)†	93.16	-	87.14	94.07	-	**78.84**	-	37.36	-	67.88
(Ping et al., 2023)	92.65	-	87.02	-	-	73.40	-	38.00	-	66.06
(Wang et al., 2022a)♣†	93.00	80.80	86.90	93.30	90.70	78.30	-	-	90.00	66.80
(Wang et al., 2023b)♣	92.94	74.71	86.66	90.47	-	78.48	-	**45.15**	-	-
GCIE w/o SC (ours)	89.05	76.90	86.24	91.26	90.66	74.10	66.70	38.19	86.90	58.64
GCIE w/o F (ours)	93.20	80.68	-	-	90.33	76.50	67.79	39.22	90.15	67.48
GCIE-unify (ours)	92.83	78.57	85.98	93.55	90.17	76.58	69.28	42.31	89.66	66.19
GCIE (ours)	**94.28**	81.15	**88.36**	**94.08**	**90.92**	77.19	**69.47**	39.54	**91.35**	**68.35**

Table 1: Overall results of GCIE on NER, RETriplet and NER&RE tasks. We report the average F1 scores on 3 random seeds. †: These models have additional training processes such as structure pre-training. ♣: The training parameters of these models (typically exceeding 10B) are an order of magnitude larger at least than that of other models. Task-specific models (upper part of the table) and unified models (lower part of the table) are separated with horizontal line.

Expert.

Although training models based on gold label can avoid the expensive cost of LLM inference, it potentially leads to inconsistency of data distribution between train stage and test stage. Unless otherwise stated, all training process is based on Recognizer prediction rather than gold label. A more critical problem in pipeline IE models is considered as error propagation. Different from inter-task pipeline model, GCIE is a intra-task pipeline framework. Error propagation of GCIE resulting in decreased generalization drive in its over-dependency on type prompts from Recognizer prediction. Combined with Anti-interference experiments, fine-tuned models are more vulnerable from prompt omission rather than redundance. For this issue, we introduce Self-Correction learning strategy to rectify the shortcoming of over-dependency. Specifically, we set a reject probability subject to Bernoulli distribution, denoted by $P_r \sim Bernoulli(\alpha_r)$ each predefined type over the whole data distribution. The value of α_r is determined by the recall score of Recognizer on development set. If a type is not predicted by Recognizer, it will be eliminated with its reject probability from the Expert input prompt. Under the Self-Correction mechanism, the originally determined type prompts have become uncertain:

$$\mathcal{P}(p_i|x) = R_i + (1 - R_i) \cdot (1 - P_{ri}) \tag{5}$$

where $\mathcal{P}(\cdot)$ is the probability that a type is selected. For a data distribution, p_i is the i-th type and x is input sentence. $R_i \in \{0, 1\}$ is the output result of the gold label.

In this way, the original type prompt tokens p are replaced by $\tilde{p} = \{\tilde{p}_1, \tilde{p}_2, ..., \tilde{p}_{|p|}\}$, which is closer to the prompt of Recognizer prediction on the test stage. Notably, $\mathcal{D}$ is replaced with $\tilde{\mathcal{D}}$ contain any negative sample, which harmfully causes model to deviate from the right optimization direction. The final optimization objective for expert learning is:

$$\mathcal{L} = \sum_{(\tilde{x},\tilde{y}) \in \tilde{\mathcal{D}}} -logp(\tilde{y}|\tilde{x}, \tilde{p}; \theta) \tag{6}$$

4 Experiments

To validate the efficacy of the proposed methodology and delve into certain pivotal factors within the GCIE framework, we systematically conducted an extensive series of experiments, which includes performance evaluation of GCIE on both supervised and few-shot settings, experiments on type recognition of Recognizer across various IE subtasks, and exploratory experiments on training strategies of Expert. For all experiments, by default, the base model of Expert is Flan-T5-large (Shen et al., 2023) and LLM is Claude2 [1]. The detail configuration on various datasets can be found in the Appendix C.

[1]https://claude.ai/

Model	ED	EE				ABSA			
	ACE05-Evt	ACE05-Evt		CASIE		14-res	14-lap	15-res	16-res
	Tri	Tri	Arg	Tri	Arg	Sentiment Triplet			
(Deng et al., 2021)	77.29	-	-	-	-	-	-	-	-
(Lu et al., 2021)	-	71.90	53.80	-	-	-	-	-	-
(Wang et al., 2022b)	-	73.60	55.10	-	-	-	-	-	-
(Mao et al., 2022)	-	-	-	-	-	75.52	65.27	65.88	73.67
(Lu et al., 2022)†	-	73.36	54.79	69.33	61.30	74.52	63.88	67.15	75.07
(Lou et al., 2023)†	-	72.41	55.83	71.73	63.26	**77.26**	65.51	**69.86**	78.25
(Ping et al., 2023)	-	74.08	53.92	71.46	62.91	74.77	65.23	68.58	76.02
(Wang et al., 2022a)♣†	-	69.80	52.50	-	-	-	-	-	-
(Wang et al., 2023b)♣	-	77.13	**72.94**	67.80	63.53	-	-	-	-
GCIE w/o SC (ours)	80.08	82.46	53.71	73.57	61.55	75.29	64.22	67.07	76.28
GCIE w/o F (ours)	82.62	84.37	65.98	-	-	-	-	-	-
GCIE-unify (ours)	-	84.46	64.77	71.67	63.84	-	-	-	-
GCIE (ours)	**85.54**	**84.53**	66.79	**74.40**	**65.82**	76.51	**66.48**	69.59	**79.77**

Table 2: Overall results of GCIE on ED, EE and ABSA&RE tasks. We report the average F1 scores on 3 random seeds. †: These models have additional training processes such as structure pretraining. ♣: The training parameters of these models (typically exceeding 10B) are an order of magnitude larger at least than that of other models. Task-specific models (upper part of the table) and unified models (lower part of the table) are separated with horizontal line.

Dataset		Flan-T5		Expert		GCIE	
CoNLL03	Ent	28.3	53.2	36.6	58.6	**45.2**	**74.6**
CoNLL04	Rel	16.6	52.0	21.4	56.8	**25.7**	**57.5**
ERE	Tri	21.3	46.0	20.7	48.6	**35.5**	**53.7**
ACE05-Evt	Arg	9.6	31.6	12.8	36.5	**35.3**	**54.5**
15-res	Sen	15.7	35.7	12.3	35.5	**18.4**	**41.9**
16-res	Sen	**17.6**	41.3	12.5	39.7	16.2	**48.7**

Table 3: Overall results of GCIE and baselines on few-shot settings.

4.1 Experimental Settings

Classification. Our experiments can be broadly categorized into two main parts. The first part involves the evaluation of the performance of the GCIE framework. We have selected 6 representative IE subtasks : named entity recognition (NER), joint entity and relation extraction (NER&RE), relation triple extraction (RETriplet), aspect-based sentiment analysis (ABSA), event detection (ED), and event extraction (EE). A comprehensive performance evaluation of GCIE and its three variants (without Filtering, Self-Correction and Unifying) has been conducted for them under both supervised and few-shot settings. The second part is a discussion of the type recognition capabilities of the Recognizer in GCIE. We chose Claude2 as the LLM to conduct comparative experiments on its abilities in type recognition and standard settings, as well as its ability to recognize negative samples.

Datasets. In our experiments, all datasets used in supervised and few-shot settings includes CoNLL03 (Sang and Meulder, 2003), GENIA (Kim et al., 2003), CoNLL04 (Roth and Yih, 2004), SciERC (Luan et al., 2018), NYT (Riedel et al., 2010), ERE (Song et al., 2015), ACE05 (Christopher Walker, 2006), CASIE (Satyapanich et al., 2020), SemEval-14 (Pontiki et al., 2014), SemEval-15 (Pontiki et al., 2015), SemEval-16 (Pontiki et al., 2016). For the aforementioned datasets, both the preprocessing and evaluation for specific IE task followed previous research works, which can be found in the first part of our experiments.

4.2 Experiments on GCIE

4.2.1 Supervised Settings

Main results of the performance evaluation of GCIE on supervised settings are shown as table 1 and table 2. Among them, GCIE-unify represents the unified model on all datasets, SC represents the self-correction learning mechanism, and F represents the negative sample filtering mechanism. GCIE achieves quite impressive scores on all datasets of 6 IE tasks. Especially in certain subtasks, such as NER, EE and ABSA, our GCIE outperforms all other models, which include task-specific models and unified models. Additionally, we attempt to maintain a unified set of parameters

Dataset	Element	n	Roberta-large			Claude2 k=2			Claude2 k=5			Claude2 k=10		
			P	R	F	P	R	F	P	R	F	P	R	F
CoNLL03	Ent	30	92.6	90.8	91.7	87.7	91.5	89.6	91.3	96.3	93.7	93.4	98.6	**95.9**
SciERC	Ent	30	70.2	63.3	66.6	61.7	67.6	64.5	71.4	83.3	**76.9**	-	-	-
ACE05-Rel	Ent	40	88.6	84.8	-	76.5	90.8	83.0	78.6	94.2	85.7	82.6	96.4	**89.0**
CoNLL04	Ent	40	84.7	87.1	85.9	86.4	91.8	89.0	90.6	98.0	94.2	93.4	98.0	**95.6**
	Rel	30	79.4	77.0	78.2	76.9	84.8	80.7	80.0	90.6	**85.0**	-	-	-
ACE05-Evt	Evt	100	86.7	82.3	84.4	86.5	91.8	89.1	88.1	96.4	**92.1**	-	-	-
	Arg	80	69.0	63.3	66.0	67.6	75.0	71.1	73.3	83.3	**78.0**	-	-	-
14-res	Sen	30	87.5	87.0	87.2	81.4	91.5	86.2	81.6	93.1	87.0	89.2	95.5	**92.2**
14-lap	Sen	30	89.2	83.7	86.4	79.8	94.0	86.3	79.5	96.3	87.1	84.1	98.1	**90.6**

Table 4: The results of type recognition of Roberta and Claude2 on the dev sets of various datasets. Roberta-large is fine-tuned on full-sample train set for each datatset. n is the maximum number of examples.

for all IE tasks (GCIE-unify). In this case, we observe a slight decrease in model performance across all datasets, but it still remain close to state-of-the-art IE models. We list the important conclusions and analysis from our experiments as follows:

(1) GCIE achieves the excellent performance compared to, even exceed state of the art IE models. In most cases, our model competes with its counterparts with much fewer training parameters, which is beneficial from collaboration of LLM and SLM in negative sample filtering, type recognition and Self-Correction learning mechanism.

(2) The performance of GCIE varies significantly across different tasks, with a notable improvement in event extraction on ACE05-Evt compared to other models. This is due to the inherent sparsity and clear long-tail distribution among events of the ACE05-Evt dataset, making the model prone to overfitting. Our Recognizer, on the other hand, exhibits remarkable accuracy in event and argument type recognition. Therefore, with proper guidance, Expert can effectively learn the mapping between input and output.

(3) The Self-Correction learning mechanism is capable of correcting the expert model's reliance on type indications, and omitting it would result in a performance decline.

(4) Due to the lack of uniformity in type definitions and span granularity, arbitrarily mixing multiple dataset as training data would lead to a noticeable performance decline.

4.2.2 Few-shot Settings

To explore the performance of GCIE in resource-constrained scenarios, we get random samples from train set on 1 shot and 10 shot settings respectively for each IE task, and evaluate on the full-sample test set. We repeat 10 times for every experiment and apply the same evaluation metrics with supervised settings. Without type indicating from Recognizer, Expert instead utilize SSI and SEL proposed by UIE. Flan-T5 is the base model of Expert with complete type prompt. As shown in Table 3, GCIE outperforms both Flan-T5 and Expert by a significant margin across all datasets. We observed that, especially in some complex structured tasks (such as event extraction), both prompt-based Flan-T5 and Expert fail to learn the input-to-output mapping effectively when type indications are lacking, and they completely lack the ability to discern negative samples. Instead, the Recognizer in GCIE requires only a few examples to have sufficient capability to identify potential types and negative samples.

4.3 Experiments on Recognizer

The overall performance of GCIE heavily relies on the accuracy of Recognizer in type recognition. To investigate the effectiveness and applicability of Recognizer, we design a unified type recognition task for all IE tasks, with the aim of determining whether predefined types exist in a given text. We treat type recognition as a multi-label classification task and use F1 score as its evaluation metric.

Due to rather differences in structures and objectives among various IE subtasks, we design unique instructions (Detailed information can be found in the appendix C) as prompts of Claude2 for each IE subtask. Each instruction contains a few examples, which are regarded as primary hyperparameters in Recognizer. In addition, we fine-tune a Roberta (Liu et al., 2019) as basedline for every dataset.

As shown in Table 4, with the number of examples increasing, the performance of Claude2 shows a stable upward trend. Due to the limitations of LLM on input length, our set a maximum of in-

struction for each dataset. As the number of examples increases, Claude2's performance surpass that of fine-tuned Roberta-large rapidly on all IE subtasks, especially difficult tasks like event extraction. The most surprising point is that Claude2 has much greater recall scores than their precision over all datasets, which implies the LLM solves the problem for our instructions with high confidence. Additionally, we can draw the following conclusions:
(1) Compared to smaller models, Claude2 has stronger robustness and generalization abilities. It has a smaller variance in its scores to different datasets and performs admirably on more challenging datasets.
(2) LLM as a type recognizer has low reliance on labeled data. It achieves excellent results with just a few examples on many datasets. In practice, it's rational to take advantage of the LLM property of high recalls to guide SLM extraction.

In most case, although large language model is not a good few-shot information extractor, but is good as a type recognizer, which filters out the vast majority of negative samples and indicates Expert to extract corresponding spans.

5 Related Work

Our work involve a series of topic of NLP field including information extraction, instruction fine-tune on pre-trained language model, few-shot learning, prompt-based large language model, structure generation, etc. From the perspective of the target tasks, we mainly present research works with different paradigms for IE. Many works focus on single specific IE task, such as entity and relation extraction (Shen et al., 2022; Li et al., 2022; Yan et al., 2021; Tang et al., 2022; Shen et al., 2021; Zhong and Chen, 2020; Cui et al., 2021; Shang et al., 2022; Wei et al., 2020; Souza et al., 2019; Ye et al., 2022; Wang et al., 2020), event detection and argument extraction (Liu et al., 2023; Wang et al., 2023a; Zhang et al., 2022; Deng et al., 2021; Liu et al., 2018; Sheng et al., 2021; Lu et al., 2021; Xu et al., 2021b; Wang et al., 2022c) and aspect-based sentiment analysis (Xu et al., 2021a; Li et al., 2023, 2021; Zhou et al., 2020; Liang et al., 2022; Wu et al., 2020; Xu et al., 2020; Mao et al., 2022). Some of these works are based on few-shot settings.

With the development of community of deep language models and information extraction, an increasing number of models are designed to adopt a unified paradigm to address various different IE tasks. Early unified paradigm models typically employ multi-task joint training to enable the model to adapt to various information extraction tasks with different objectives and schema (Luan et al., 2019; Wadden et al., 2019; Lin et al., 2020). And (Lou et al., 2023) has utilized unified semantic matching to achieve state-of-the-art performance on multiple datasets. The most recent research effort (Peng et al., 2023; Ping et al., 2023; Gao et al., 2023) aim to introduce novel method to adapt universal IE tasks rather than unified modeling. However, the most closely related approach to our work is still the unified structured generation paradigm for a range of IE tasks (Lu et al., 2022; Wang et al., 2022a, 2023b). In the era of LLMs, how to eliminate the vast gap of LLM in IE tasks and further achieve new SOTA performances, perhaps become the hottest research direction currently (Gui et al., 2023; Wang et al., 2023c; Wadhwa et al., 2023b).

6 Conclusion

In this study, We analyze the important factors influencing the performance of the IE models and introduce three core capabilities, which usually cannot be possessed simultaneously by existing IE models. Through prompt design and a series of exploration experiments, we find that in-context based LLM is able to identify negative samples and recognize predefined type information from texts. Based on this, We propose GCIE, which combines the strengths of LLM and Experts in IE tasks to encompass both of these capabilities. With the LLM-based Recognizer and the unified structured generation Paradigm-based Expert, GCIE is designed to be general for all IE tasks. Extensive experiments confirm that, compared to existing LLM-only and SLM-only methods, GCIE can further enhance performance on all IE tasks effectively. Furthermore, we explore the impact of demonstration format and label type format on in-context learning and supervised fine-tuning. All of these indicate a prospective unified IE research direction to take advantages of LLM and fine-tuned SLM.

Limitations

Despite the outstanding performance achieved by our approach, some obvious limitations should be pointed out and addressed in the future: 1) Additional inference latency brought by LLM compared to SLM-only methods. 2) Task-specific prompts that require carefully crafted manual design for LLM. 3) Sensitive hyperparameters settings in self-correction mechanism.

References

Peggy M. Andersen, Philip J. Hayes, Alison K. Huettner, Linda M. Schmandt, Irene B. Nirenburg, and Steven P. Weinstein. 1992. Automatic extraction of facts from press releases to generate news stories. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, page 170–177, USA. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Julie Medero Kazuaki Maeda Christopher Walker, Stephanie Strassel. 2006. Ace 2005 multilingual training corpus.

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, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1835–1845. Association for Computational Linguistics.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Tou Huaixiao, Mosha Chen, Fei Huang, and Huajun Chen. 2021. OntoED: Low-resource event detection with ontology embedding. 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 2828–2839, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Manqing Dong, Chunguang Pan, and Zhipeng Luo. 2021. Mapre: An effective semantic mapping approach for low-resource relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2694–2704. Association for Computational Linguistics.

Chang Gao, Wenxuan Zhang, Wai Lam, and Lidong Bing. 2023. Easy-to-hard learning for information extraction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11913–11930. Association for Computational Linguistics.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, page 677–692.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Honghao Gui, Jintian Zhang, Hongbin Ye, and Ningyu Zhang. 2023. Instructie: A chinese instruction-based information extraction dataset. *CoRR*, abs/2305.11527.

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. *CoRR*, abs/2305.14450.

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. *CoRR*, abs/2012.14978.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics.

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, June 29 - July 3, 2003, Brisbane, Australia, pages 180–182.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10965–10973. AAAI Press.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard H. Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6319–6329. Association for Computational Linguistics.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 6132–6142. ACM.

Shuo Liang, Wei Wei, Xian-Ling Mao, Fei Wang, and Zhiyong He. 2022. BiSyn-GAT: Bi-syntax aware graph attention network for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7999–8009. Association for Computational Linguistics.

Jian Liu, Dianbo Sui, Kang Liu, Haoyan Liu, and Zhe Zhao. 2023. Learning with partial annotations for event detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 508–523. Association for Computational Linguistics.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13318–13326. AAAI Press.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2795–2806. Association for Computational Linguistics.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5755–5772. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 3036–3046. Association for Computational Linguistics.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1956–1971. Association for Computational Linguistics.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Ion Muslea. 1999. Extraction patterns for information extraction tasks: A survey.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan T. McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Trans. Assoc. Comput. Linguistics*, 9:1475–1492.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2023. Bidirectional language models are also few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tianshuo Peng, Zuchao Li, Lefei Zhang, Bo Du, and Hai Zhao. 2023. FSUIE: A novel fuzzy span mechanism for universal information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16318–16333. Association for Computational Linguistics.

Yang Ping, Junyu Lu, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Pingjian Zhang, and Jiaxing Zhang. 2023. Uniex: An effective and efficient framework for unified information extraction via a span-extractive perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16424–16440. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

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, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer.

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 (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Erik F. 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, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

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, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8749–8757. AAAI Press.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6311–6322. Association for Computational Linguistics.

Yuming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11285–11293. AAAI Press.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705.

Yongliang Shen, Xinyin Ma, Yechun Tang, and Weiming Lu. 2021. A trigger-sense memory flow framework for joint entity and relation extraction. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1704–1715. ACM / IW3C2.

Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. Parallel instance query network for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 947–961. Association for Computational Linguistics.

Jiawei Sheng, Shu Guo, Bowen Yu, Qian Li, Yiming Hei, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2021. Casee: A joint learning framework with cascade decoding for overlapping event extraction. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 164–174. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. Unirel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7087–7099. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5783–5788. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023a. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023b. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

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, Dublin, Ireland, May 22-27,* *2022*, pages 803–823. Association for Computational Linguistics.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022b. Query and extract: Refining event extraction as type-oriented binary decoding. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 169–182. Association for Computational Linguistics.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022c. Query and extract: Refining event extraction as type-oriented binary decoding. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 169–182. Association for Computational Linguistics.

Sijia Wang, Mo Yu, and Lifu Huang. 2023a. The art of prompting: Event detection based on type specific prompts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1286–1299. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *CoRR*, abs/2304.08085.

Xingyao Wang, Sha Li, and Heng Ji. 2023c. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3640–3663. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1572–1582. International Committee on Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1476–1488. Association for Computational Linguistics.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *CoRR*, abs/2010.04640.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021a. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4755–4766. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2339–2349. Association for Computational Linguistics.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021b. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3533–3546. Association for Computational Linguistics.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 185–197. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4904–4917. Association for Computational Linguistics.

Hongming Zhang, Wenlin Yao, and Dong Yu. 2022. Efficient zero-shot event extraction with context-definition alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7169–7179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for joint entity and relation extraction. *CoRR*, abs/2010.12812.

Jie Zhou, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2020. SK-GCN: modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. *Knowl. Based Syst.*, 205:106292.

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

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2450–2460. ACM.

A Task and Dataset

In this study, we conducted experiments on 6 information extraction tasks and 13 datasets. We provide a detailed description of each task, dataset, and evaluation method as follows. The detail statistics of all IE dataset we use are shown in Table 5.

Named Entity Recognition is a task in natural language processing that focuses on identifying and classifying named entities mentioned in text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. An entity mention is correct if its offsets and type match a reference entity.

Relation Triplet Extraction is a task in NLP that aims to identify and extract structured information from text by identifying relationships between entities mentioned in the text. An RTE system typically takes as input a sentence or a document and outputs a set of triples, where each triple consists of a subject entity, a relation, and an object entity. A relation triplet is correct if its relation type is correct and the string of the subject/object are correct.

Joint Entity and Relation Extraction is a task that aims to identify and extract entities and their relations from textual data. It involves the identification of both entities (e.g., people, places, organizations) and the relationships that exist between these entities within a text. A relation is correct if its relation type is correct and the offsets and entity types of the related entity mentions are correct.

Event Detection is a task in natural language processing that aims to identify and extract key informational elements from text, which are known as 'events'. These events are semantic units marked by a trigger phrase in text that describe meaningful occurrences or actions within a text. A event is correct if its trigger offsets and type match a reference trigger.

Dataset	Elements	Sentences		
		Train	Dev	Test
CoNLL03	4 Ent	14,041	3,250	3,453
GENIA	5 Ent	15,038	1,654	1,854
ACE05-Ent	7 Ent	7,299	971	1,060
NYT	1 Ent, 24 Rel	56,196	5,000	5,000
CoNLL04	4 Ent, 5 Rel	922	231	288
SciERC	6 Ent, 7 Rel	1,861	275	551
ACE05-Rel	7 Ent, 6 Rel	10,051	2,420	2,050
ERE	38 Evt	13,736	1,000	1,163
ACE05-Evt	33 Evt, 22 Arg	19,240	902	676
CASIE	5 Evt, 13 Arg	11,189	1,778	3,208
14res	1 Asp, 3 Sen	1,266	310	492
14lap	1 Asp, 3 Sen	906	219	328
15res	1 Asp, 3 Sen	605	148	322
16res	1 Asp, 3 Sen	857	210	326

Table 5: Statistics of all IE dataset in this study.

Event Extraction is a task that aims to identify and extract key information about events from textual data. These events can be any significant occurrence or transaction, such as accidents, attacks, elections, or births. It is typically decomposed into two sub-tasks: event trigger detection and event argument extraction, which can be performed either in a pipeline or an end-to-end manner. An event trigger is correct if its offsets and event type matches a reference trigger. An event argument is correct if its offsets, role type, and event type match a reference argument mention.

Aspect-based Sentiment Analysis is a subtask of sentiment analysis, which aims to identify the sentiment expressed in text towards specific aspects of an entity, such as a product, service, or event. ABSA often involves two primary tasks: aspect and opinion extraction and aspect sentiment classification. A sentiment triplet consists of an aspect, an opinion and their sentiment polarity. A correct triplet requires the offsets boundary of the target, the offsets boundary of the opinion span, and the target sentiment polarity to be all correct at the same time.

B Anti-interference Test

Negative samples those are scarcely informative or lacking of demand-oriented annotation commonly appear in the realm of information extraction. In this study, we investigate the impact of negative samples on model performance. A series of experiments indicate negative recognition is a pivotal ability to conduct practical IE tasks. Specifically, we fine-tune small language model with structural

Dataset	Recognizer	Expert	Supervised				Few-Shot			
			batch	learning rate	label smoothing	examples	batch	learning rate	label smoothing	examples
CoNLL03			16	5e-5	0	30	8	5e-5	0	4, 30
GENIA			16	5e-5	0	30	-	-	-	-
ACE05-Ent			16	5e-5	0	40	-	-	-	-
NYT			16	5e-5	0	75	8	5e-5	0	25, 75
CoNLL04			16	5e-5	0	50	4	5e-5	0.1	10, 50
SciERC			8	5e-5	0.1	70	4	5e-5	0.1	14, 70
ACE05-Rel	Claude2	Flan-T5-large	8	5e-5	0.1	70	4	5e-5	0.1	14, 70
ERE			-	-	-	-	8	5e-5	0	5, 80
ACE05-Evt			16	5e-5	0.1	100	8	5e-5	0.1	34, 100
CASIE			16	5e-5	0.1	72	8	5e-5	0.1	18, 72
14res			16	5e-5	0.05	15	-	-	-	-
14lap			16	5e-5	0.05	15	-	-	-	-
15res			16	5e-5	0.05	15	8	5e-5	0.1	3, 15
16res			16	5e-5	0.05	15	8	5e-5	0.1	3, 15

Table 6: Hyper-parameters for GCIE training on both supervised and few-shot settings.

generative paradigm on ACE05-Evt dataset to describe the variation trend of model performance, by scaling the proportion of negative samples in the total training numbers, shown as Figure 3. From the result, it is clear that a high proportion of filtration is beneficial to predicting positive samples and harmful to recognizing negative samples. we attribute this phenomenon to model overfitting on certain data distribution explained by a example (see Figure 4). Additionally, according to the results of "self" curve, when the number of negative samples is reduced to a certain extent, the simulated performance tends to be similar to the gold performance. To some extent, negative samples simultaneously enhance the robustness of a fine-tuned model with limited data and weaken its ability of valid information identification. It is plausibly ideal to correctly identify negative samples without parameter variation.

One step further, we investigate the capacities of negative sample recognition based on prompt-based LLM and fine-tuned SLM. As seen in Table 7, we compute the accuracy on development sets across three IE dataset. In comparison to SLM, LLM with few examples seems exhibit powerful talent on negative sample recognition, with a much great margin. On the basis of the examination, we select LLM as negative sample filter to implicitly improve the robustness of our IE system. And more effective ways remain more endeavors in our follow-up research works.

C Experiment Details

In this section, we describe details of experiments that include hyper-parameters on supervised and few-shot settings, Recognizer prompt construction and Expert prompt construction.

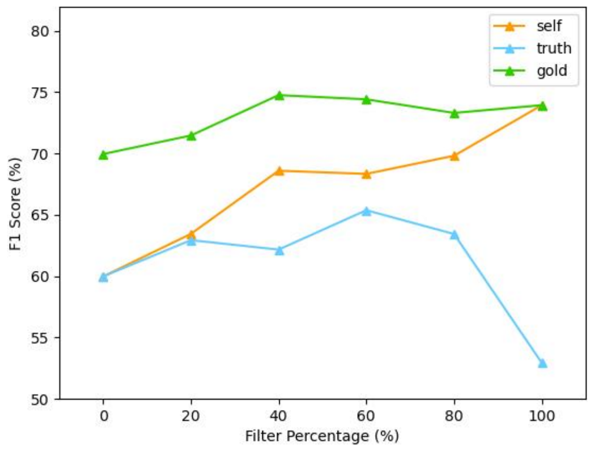

Figure 3: The performance on ACE05-Evt of generative fine-tuned models with negative sample filteration in varying proportions. "self" denotes the score on simulated labels by random sampling at the ratio; "truth" denotes the score on practical labels; "gold" denotes the score on positive labels.

C.1 Hyper-parameters

As shown in Table 6, on supervised and few-shot experiments, we select Claude2 and Flan-T5-large as LLM and fine-tune base model, AdamW (Loshchilov and Hutter, 2019) as optimizer with learning rate=5e-5 for all dataset. Label Smoothing (Szegedy et al., 2016) are applied for partial IE tasks to alleviate overfitting. To accomplish all our experiments successfully, a 48G memory are required at least.

C.2 Recognizer Prompt

We manually design unique instructions for each dataset, which can be divided into two parts: task descriptions and reference demonstrations. The task description part explains to LLM the task we are conducting and predefined label types. The

Model	ACE05-Evt	SciERC	CoNLL03
UIE-SEL p=1.0	83.9	75.0	93.1
UIE-SEL p=0.6	81.9	68.8	92.7
UIE-SEL p=0.2	77.8	50.0	91.5
Claude2 k=5	85.5	75.0	95.0

Table 7: The results of negative sample recognition based on few-shot LLM and fine-tuned SLM. p is the proportion of negative samples used in training.

reference demonstrations part includes samples selected from the training set, which are processed into input-output pairs. The performance of in-context learning of LLM can be improved by outputs designed with Chain-of-Thought (Wei et al., 2022). As shown below, the large model analyzes the input text according to our instructions and generates output in the same format as the examples.

CoNLL03

We are conducting named entity recognition task.
We only consider three entity types: Person(a specific person name), Organization(an specific organization) and Location(a specific place). Please note that a sentence probably does not contain any defined entity.
There are several pairs of input and output as examples.
sentence: EU rejects German call to boycott British lamb .
result: [Organization : EU] [Person : none] [Location : none]
sentence: The guitarist died of a drugs overdose in 1970 aged 27 .
result: [Person : none] [Organization : none] [Location : none]
sentence: China says Taiwan spoils atmosphere for talks .
result: [Location : China] [Location : Taiwan] [Person : none] [Organization : none]
sentence: BEIJING 1996-08-22
result: [Location : BEIJING] [Person : none] [Organization : none]
......
Let us start! Please analyse the following sentence and complete the result.
sentence: He was well backed by England hopeful Mark Butcher who made 70 as Surrey closed on 429 for seven , a lead of 234 .
result:

GENIA

We are conducting named entity recognition task.
entity types are defined as follows:
1. Protein : the name of certain protein.
2. DNA : the name of certain DNA.
3. RNA : the name of certain RNA.
4. Cell line : the name of certain cell line.
5. Cell type : the name of certain cell type.
There are several pairs of input and output as examples.
sentence: Thyroid hormone receptors form distinct nuclear protein- dependent and independent complexes with a thyroid hormone response element .
result: [Protein : Thyroid hormone receptors] [DNA : thyroid hormone response element] [RNA : none] [Cell line : none] [Cell type : none]
sentence: TR alpha 1 and TR beta 2 each formed a single major TR : TREp complex which comigrated with the least retarded complex formed by GH3 NE , while TR beta 1 formed multiple complexes suggesting that it can bind to TREp as an oligomer .
result: [Protein : TR alpha 1] [Protein : TR beta 2] [DNA : none] [RNA : none] [Cell line : none] [Cell type : none]
sentence: Human immunodeficiency virus type 1 (HIV-1) can establish a persistent and latent infection in CD4+ T lymphocytes (W.C.Greene , N.Engl.J. Med.324 : 308-317 , 1991 ; S.M.Schnittman , M.C.Psallidopoulos , H.C. Lane , L.Thompson , M.Baseler , F.Massari , C.H.Fox , N.P.Salzman , and A.S.Fauci , Science 245 : 305-308 , 1989) .
result: [Protein : none] [DNA : none] [RNA : none] [Cell line : none] [Cell type : CD4+ T lymphocytes]
sentence: Such changes clearly can not be explained by genomic mechanisms , which are responsible for later effects than the membrane related rapid responses .
result: [Protein : none] [DNA : none] [RNA : none] [Cell line : none] [Cell type : none]
......
Let us start! Read the text and complete content of the result. Please note that a sentence probably does not contain any defined entity.
sentence: The values of plasma aldosterone and 18-OH-B were also low .
result:

NYT

We are conducting relation triplet extraction task.
entity types are defined as follows:
location, organization, person.
relation types are defined as follows:
1. (location) is the administrative divisions of (location)
2. (person) is the advisors of (person)
3. (location) is the capital of (location)
4. (person) is the children of (person)
5. (person) work for (organization)
6. (location) contains (location)
7. (location) is the place of (location)
8. (organization) is the ethnicity of (person)
9. (organization) is founded by (person)
10. (location) is distributed in (location)
11. industry
12. (organization) is located in (location)
13. (person) is a major shareholder of (organization)
14. (organization) has major shareholders with (person)
15. the nationality of (person) is (location)
16. (person) is the neighborhood of (person)
17. people
18. (organization) is founded in (location)
19. (person) live in (location)
20. (person) is born in (location)
21. (person) is died in (location)
22. (person) have a professional job in (organization)
23. (person) believes in (organization)
24. (organization) is a team in (location)
Please determine if there exist entities and relations predefined above in the given sentence.
There are several pairs of input and output as examples.
sentence: Prosecutors ' interest in Chubb may indicate that the insurance scandal is widening , even after more than a year of intense scrutiny by Eliot Spitzer , the New York attorney general , and officials at the Securities and Exchange Commission .
result: [Carolina contains Greensboro]
sentence: The historic city of Oaxaca has long been one of the most popular tourist destinations in Mexico .
result: [Oaxaca is the administrative divisions of Mexico] [Mexico is the country of Oaxaca]
sentence: They needed to beat the Red Sox , and they also needed the Chicago White Sox to beat the Cleveland Indians – which Chicago did , 4-3 .
result: [Sox is located in Chicago] [Sox is a team in Chicago]
sentence: Today , Maimonides stands for an austerely intellectual doctrinal Judaism , the castigation of all forms of idolatry and the combining of Jewish learning with secular science and philosophy -LRB- in his own times , this meant Aristotle -RRB- .
result: [Maimonides believes in Judaism]
......
Let us start! Read the text and complete content of the result.
sentence: At a conference on Sunday in Manchester in northern England , Mr. Blair 's measures drew a sharp response from some participants , including Yvonne Ridley , a former newspaper journalist in Britain who converted to Islam after being imprisoned by the Taliban in Afghanistan .
result:

SciERC

We are going to conduct named entity recognition task.
Entity types are defined as follow:
1. Task: specific academic task, application, problem to solve, such as "information extraction", "machine reading systems", "image segmentation", etc.
2. Material: data, dataset, resource, corpora, knowledge base.
3. Method: specific method, model, system, such as "language models", "CORENLP, POS profilers", "kernel methods", etc.
4. Metric: evaluation metric, such as "accuracy", "recall" and etc.
5. Generic: general term, noun, such as "approach", "method", "algorithm" and etc.
6. OtherScientificTerm: other scientific terminology.
Here are some pairs of sentence and result as examples.
sentence: This new algorithm deviates from the traditional approach of wall building and layering .
result: [Generic : algorithm] [Method : approach of wall building and layering] [Task : N/A] [Material : N/A] [OtherScientificTerm : N/A]
sentence: Graph unification remains the most expensive part of unification-based grammar parsing .
result: [Task : Graph unification] [Task :

unification-based grammar parsing] [Material : N/A] [Method : N/A] [Metric : N/A] [Generic : N/A] [OtherScientificTerm : N/A]
sentence: This task involves two core technologies : natural language processing -LRB- NLP -RRB- and information extraction -LRB- IE -RRB- .
result: [Generic : task] [Method : natural language processing -LRB- NLP -RRB-] [Task : information extraction -LRB- IE -RRB-] [Material : N/A] [Metric : N/A] [OtherScientificTerm : N/A]
sentence: Tokens are computed via a small-to-large scale grouping procedure employing a greedy , best-first , strategy for choosing the support of new tokens .
result: [Method : small-to-large scale grouping procedure] [Task : N/A] [Material : N/A] [Metric : N/A] [Generic : N/A] [OtherScientificTerm : N/A]
......
Let us start! Please analyse the following sentence and complete the result.
sentence: Holistically , a video has its inherent structure – the correlations among video frames .
result:

ACE05-Ent / ACE05-Rel

We are going to conduct named entity recognition task.
Entity types are defined as follow:
1. Person: person name, group name, personal pronoun and etc.
2. Organization: government, business, institution, association, political party and etc.
3. GPE: continent, nation, country, state, province, district, country group and etc.
4. Location: a place or area such "world", "earth", "sea", "desert" and etc.
5. Facility: a building such as "airport", "office", "restaurant", "school" and etc.
6. Vehicle: vehicle.
7. Weapon: weapon.
Here are some pairs of sentence and result as examples.
sentence: sharon spit on tab and called her names .
result: [Person : sharon] [Person : tab] [Person : her]
sentence: a spokesman says that if any charges are filed , they will be on the low end of the misdemeanor scale .
result: [Person : spokesman]
sentence: BEIJING (AP)
result: [Organization : AP] [GPE : BEIJING]
sentence: The islands are in the Yellow Sea , between the northeastern province of Liaoning and North Korea .
result: [GPE : Liaoning] [GPE : province] [GPE : North Korea] [Location : islands] [Location : Yellow Sea]
......
Let us start! Please analyse the following sentence and complete the result. sentence: That 's why you played a four-loss team for your conference title this year .
result:

ACE05-Evt

We are going to conduct event extraction task.
event types and event arguments are defined as follows:
1. type: Birth; arguments: person, place. This event is probably triggered by born, birth and etc.
2. type: Death; arguments: agent, victim, place, instrument. This event is probably triggered by die, kill, eliminate, eradicate and etc.
3. type: Marriage; arguments: person, place. This event is probably triggered by marry, wed and etc.
4. type: Divorce; arguments: person, place. This event is probably triggered by divorce and etc.
5. type: Injury; arguments: agent, victim, place, instrument. This event is probably triggered by injure, wound and etc.
6. type: Start of position; arguments: person, affiliation, place. This event is probably triggered by hire, put, recruit, precede and etc.
7. type: End of position; arguments: person, affiliation, place. This event is probably triggered by fire, leave, retire, former, resign and etc.
8. type: Nomination; arguments: person, agent. This event is probably triggered by nominate, name, select and etc.
9. type: Election; arguments: person, affiliation, place. This event is probably triggered by elect, win, vote and etc.
10. type: Start of organization. arguments: agent, organization, place. This event is probably triggered by start, open, establish and etc.
11. type: End of organization. arguments: organization, place. This event is probably triggered by end, close and etc.
12. type: Merger. arguments: organization. This event is probably triggered by merge and etc.

13. type: Bankruptcy. arguments: organization, place. This event is probably triggered by bankrupt and etc.
14. type: Meeting. arguments: participant, place. This event is probably triggered by meet, summit, negotiate, discuss, talk and etc.
15. type: Phone contact. arguments: participant, place. This event is probably triggered by write, call, letter, phone and etc.
16. type: Transfer of ownership; arguments: buyer, seller, place, possession, beneficiary. This event is probably triggered by buy, seize, capture, sale and etc.
17. type: Transfer of money; arguments: giver, recipient, place, beneficiary. This event is probably triggered by transfer, pay and etc.
18. type: Movement; arguments: deployer, object, destination, origin, vehicle. This event is probably triggered by deploy, go, arrive, advance, land and etc.
19. type: Attack; arguments: attacker, target, victim, place, instrument. This event is probably triggered by war, force, strike, attack, fight, battle, fire, terror, hit, incident, bomb, conflict, violence, explosion, invade, kill and etc.
20. type: Demonstration; arguments: participant, place. This event is probably triggered by protest, march, rally, demonstrate and etc.
21. type: Arrest; arguments: agent, person, place. This event is probably triggered by arrest, apprehend, nab, capture and etc.
22. type: Parole; arguments: authority, person, place. This event is probably triggered by release, parole and etc.
23. type: Trial; arguments: defendant, adjudicator, prosecutor, place. This event is probably triggered by hearing, trial and etc.
24. type: Charge; arguments: defendant, adjudicator, prosecutor, place. This event is probably triggered by charge, accused, indict and etc.
25. type: Sue; arguments: plaintiff, defendant, adjudicator, place. This event is probably triggered by sue, lawsuit, suit and etc.
26. type: Convict; arguments: defendant, adjudicator, place. This event is probably triggered by convict, guilty, verdict and etc.
27. type: Sentence; arguments: defendant, adjudicator, place. This event is probably triggered by sentence, condemn, face and etc.
28. type: Fine; arguments: payor, adjudicator, place. This event is probably triggered by fine, pay and etc.
29. type: Execute; arguments: agent, person, place. This event is probably triggered by execute, kill and etc.
30. type: Extradite; arguments: agent, destination, origin. This event is probably triggered by extradite and etc.
31. type: Acquit; arguments: defendant, adjudicator. This event is probably triggered by acquit and etc.
32. type: Pardon; arguments: defendant, adjudicator, place. This event is probably triggered by pardon and etc.
33. type: Appeal; arguments: plaintiff, adjudicator, place. This event is probably triggered by appeal and etc.

You are required to infer whether any event predefined above is mentioned in a sentence and answer with the format: "[event type : trigger (argument : tokens)...(argument : tokens)]..." or "There is no event mentioned in the sentence". Events that have happened in the past, are happening now, or may occur in the future should all be taken into consideration, but those events not defined by us should be overlooked. Here are several examples.

sentence: Here are some of the fine achievements of the terrorist Marwan Barghouti Marwan Barghouti (born June 6 , 1958) is a Palestinian leader from the West Bank and a leader of the Fatah movement that forms the backbone of the Palestinian Authority and the Palestine Liberation Organization (PLO) .
result: [Birth : born (person : Marwan Barghouti) (place : West Bank)]

sentence: If you go for a home birth you can rent a birthing pool . I would n't necessaritly say that you will have a repeat labour ! My first labour I was 30 hours and had an epidural after 22 hours . I went in saying " give me the epidural asap - and never got to the state where I felt that I needed it .
result: [Birth : birth (person : unspecified) (place : unspecified)]

sentence: The birth comes days after the death of O'Neal 's maternal grandfather , Sirlester O'Neal . result: [Birth : birth (person : unspecified) (place : unspecified)] [Death : death (victim : grandfather) (agent : unspecified) (place : unspecified) (instrument : unspecified)]

sentence: Shaunie O'Neal gave birth to the couple 's third child at 1:52 a.m. at a Los Angeles - area hospital , team spokesman John Black said .

result: [Birth : birth (person : child) (place : hospital)]
sentence: police are now considering the possibility that the remains are those of laci peterson and her unborn child .
result: [Birth : unborn (person : child) (place : unspecified)]
sentence: But we should n't lose sight of the fact that we have two political parties so people will have choices .
result: There is no event mentioned in the sentence.
sentence: SANDERS Well it 's not – are you suggesting that when tens and thousands of Iraqi women and children are killed , and when young men and women in this country are unnecessarily put at harm 's risk , what should we do ?
result: [Death : killed (victim : children) (agent : unspecified) (place : unspecified) (instrument : unspecified)]
sentence: " They make this look like a John Wayne movie , " said protester Elvis Woods .
result: There is no event mentioned in the sentence.
......
Let us start! Read the following sentence and output the correct result.
sentence: He had to sue to become our president , and he keeps trying to bribe other countries ' democratic governments into his supporting his agenda .
result:

CASIE

We are conducting cybersecurity event extraction task.
event types and their optional argument roles are defined as follows:
1. Data Breach[compromised data, number of data]
2. Phishing[trusted entity]
3. Ransom[ransom price, payment method]
4. Discover Vulnerability[discoverer, capability, system owner]
5. Patch Vulnerability[releaser, issue, patch, number of patch, platform]
Please determine if there have events predefined above in the given sentence. Events that have happened in the past, are happening now, or may occur in the future should all be taken into consideration. Here are several examples.
example1
sentence: A former Chicago Public Schools worker faces several felony charges after officials allege the worker stole personal information on about 80,000 employees, volunteers and vendors from a CPS database. The former worker, Kristi Sims, was arrested Thursday...The data included children's names, home and cellphone numbers, email addresses and ID numbers.
result: There are Data Breach[Compromised Data-personal data]: stolen in the sentence.
......
Let us start! Read the sentence and complete content of the result.
sentence: The two vulnerabilities are critical remote code execution flaws that exist in Adobe Photoshop CC. Adobe hurried out unscheduled patches today for two critical flaws that could...there may have been a disclosure deadline and the release did not make this month's typical release cycle but needed to release before September's release cycle.
result:

SemEval-14 / 15 / 16

We are conducting aspect-based sentiment analysis task. What you need to do is to recognize the sentiments (positive, negative, neutral) implied in the sentence.
Here are some examples.
example1
sentence: I charge it at night and skip taking the cord with me because of the good battery life .
result: good is a positive opinion for battery life; Therefore, there have positive sentiment but no negative, neutral sentiments in the sentence.
example2
sentence: The price premium is a little much , but when you start looking at the features it is worth the added cash .
result: worth is a positive opinion for features; much is a negative opinion for price premium; Therefore, there have positive, negative sentiments but no neutral sentiment in the sentence.
example3
sentence: Until I bought the Dell , I thought you just looked for what you wanted (size , software , options , hardware) and purchase the best deal you could find .
result: best is a neutral opinion for hardware; Therefore, there have neutral sentiment but no positive, negative sentiments in the sentence.
......

Let us start! Read the sentence and complete content of the result.
sentence: We also use Paralles so we can run virtual machines of Windows XP Professional , Windows 7 Home Premium , Windows Server Enterprise 2003 , and Windows Server 2008 Enterprise .
result:

C.3 Expert Prompt

The Expert prompt for each input text is to tell language model what types exist probably in the given sentence. The results of Expert prompts drive in Recognizer but they are not required to be very meaningful to be understood by human or LLM. Underlying our observation, it is most worthy that type words ought to be designed distinctively against informative mentions. To this end, we list handcrafted Expert prompts as follow.

CoNLL03: person, organization, location, others.

CoNLL04: person, organization, location, others, based in, work for, located in, live in, kill.

SciERC: Task, Material, Method, Metric, Generic, Others, Part of, Used for, Hyponym of, Conjunction with, Feature of, Evaluate for, Compare with.

ACE05-Rel: Person, Organization, Location, Geographical political entity, Facility, Vehicle, Weapon, Physical, Part whole, Personal social, Organization affiliation, Agent artifact, General affiliation.

ACE05-Evt: Acquit, Appeal, Arrest, Attack, Born, Charge, Convict, Bankrupt, Demonstrate, Die, Elect, Divorce, End-Organization, End-Position, Execute, Extradite, Fine, Injure, Marry, Meet, Merge, Nominate, Pardon, Phone, Parole, Sentence, Start-Organization, Sue, Start-Position, Transfer-Money, Transfer-Ownership, Transport, Trial-Hearing, Vehicle, Artifact, Destination, Person, Agent, Entity, Place, Target, Attacker, Giver, Recipient, Plaintiff, Victim, Buyer, Seller, Instrument, Origin, Organization, Beneficiary, Defendant, Adjudicator, Prosecutor.

GENIA: Protein, DNA, RNA, Cell line, Cell type.

NYT: administrative division, advisor, capital, children, company, contain, country, ethnicity, founder, geographic distribution, industry, location, major shareholder of, major shareholder, nationality, neighborhood of, people, place of finding, place of living, place of birth, place of death, profession, religion, team.

CASIE: Data Breach, Phishing, Ransom, Discover Vulnerability, Patch Vulnerability, Compromised data, Number of data, Trusted entity, Ransom price, Payment method, Discoverer, Capability, System owner, Releaser, Issue, Patch, Number of patch, Platform.

SemEval-14/15/16: aspect, opinion, positive, negative, neutral.

D Further Exploration

D.1 Demonstration Format

The examples determine the quality of outputs of LLM in type recognition. Whether inputs (with prompts) and outputs of LLM should include pairs of types and mentions or only type clues? We conduct further exploration on this issue, and the results are shown in Table 8. On the SciERC, we find that explicitly providing spans result in much better performance compared to only providing types. However, on the CoNLL03 and 14-res, there are only a slight improvement. This is because the entity and relation labels in the SciERC have abstract semantics, and Claude2 needs more contextual information to understand the label semantics. Leveraging span mentions reasonably enhances the in-context learning ability of LLM, analogous to CoT in relation extraction (Wei et al., 2022; Wadhwa et al., 2023b).

D.2 Label Type Format

When Expert receives type indications as prompts and generates structured text, it treats labels as natural language phrases. This is done to fully leverage the knowledge that the language model has acquired during the pre-training phase. However, can this approach truly effectively utilize the knowledge stored in pre-trained language models? To validate this perspective, we conduct exploratory experiments on partial entity and relation labels on the CoNLL04 and SciERC datasets, the results

Model	CoNLL03	SciERC	14-res
	Ent	Rel	Sen
Claude2 w/o Span	89.8	57.2	86.1
Claude2	93.7	65.7	87.0
△Gain	+3.9	+8.5	+0.9

Table 8: Results of ablation experiments of the example format on entity, relation and sentiment.

Model	CoNLL04		SciERC	
	Loc	Located_In	Gerneric	USED-FOR
Expert-A	88.6	76.9	68.6	41.9
Expert-B	90.3	78.6	68.3	42.1
Expert-C	89.3	77.7	69.0	40.5

Table 9: Results of ablation experiments of the Type Phrase on CoNLL04 and SciERC dev sets.

of which are shown in Table 9. Expert-A treats labels as specific symbols. Expert-B uses meaningful words "place", "located in", "generic" and "used for" as type phrases. Expert-C substitutes them with abstract words such as "Located-in". The results show that the entities of "Loc" type are more susceptible to label semantics than entities of type "Generic". In contrast, relations are much less affected by label semantics.

E Schema Format

The format of outputs both used for LLM and Expert conform to structure generation shown as Figure 5. It is noted that the outputs of LLM are not responsible to generate correct values, instead it's only required to determine what defined information a sentence probably mentions. While we observe that offering a complete answer in which the values serve as explainable evidences is beneficial for LLM to think step by step.

sentence: Michael York , one of Jack Welch 's attorneys , called the move routine .
result: { [Movement : move] }
label: { }

sentence: Turkish party leader Recep Tayyip Erdogan named prime minister , may push to allow in U.S. troops .
result: { [Nomination : named] }
label: { }

Figure 4: An example about overfitting results on negative samples of a fine-tuned generative IE model.

[Person : Oswald] [Location : Mediterranean]

(a) Named Entity Recognition

[People : James Hackett (Work for : Titan Systems) (Live in : U.S)]

(b) Relation Extraction

[Material : uncalibrated images (Used for : surface re-flectance estimates)] [Method : surface re-flectance estimates

(c) Joint Entity and Relation Extraction

[Aspect : the food (Positive : decilious)] [Aspect : service (Negative : a little bad)]

(d) Aspect-based Sentiment Analysis

[End-Position : leave]

(e) Event Detection

[Meet : talks (Entity : Bush) (Place : retreat)] [Transport : arrived (Artifact : Blair) (Destination : Washington)]

(f) Event Extraction

Figure 5: There are schema examples from (a) to (f) corresponding to six information extraction tasks.