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

K Bao ✉
School of Computer Science
Fudan University

Ning Wang
School of Computer Science
Fudan University

Abstract

Recently, unified information extraction has garnered widespread attention from the NLP community, which aims to use a unified paradigm to perform various information extraction tasks. However, prevalent unified IE approaches inevitably encounter challenges such as noise interference, abstract label semantics, and diverse span granularity. In this paper, we first present three problematic assumptions regarding the capabilities of unified information extraction model. Furthermore, we propose the **G**eneral **C**ollaborative **I**nformation **E**xtraction (GCIE) framework to address these challenges in universal information extraction tasks. Specifically, GCIE consists of a general Recognizer as well as multiple task-specific Experts for recognizing predefined types and extracting spans respectively. The Recognizer is a large language model, while the Experts comprise a series of smaller language models. Together, they collaborate in a two-stage pipeline to perform unified information extraction. Extensive empirical experiments on 6 IE tasks and several datasets, validate the effectiveness and generality of our approach.

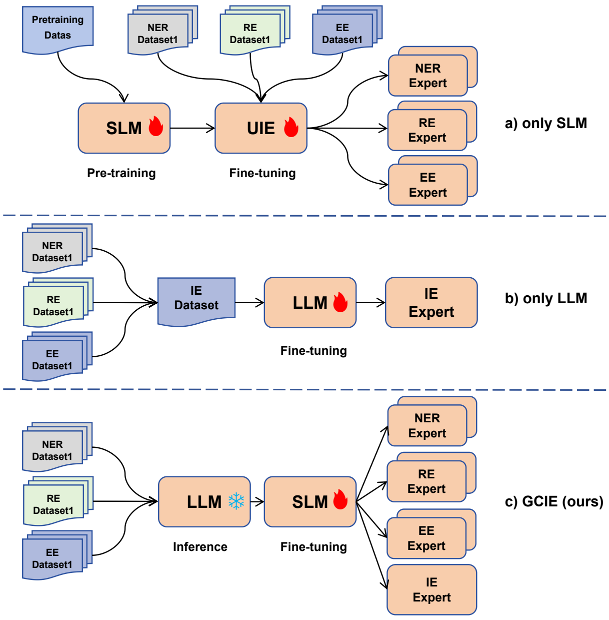

Figure 1: The paradigms of GCIE and currently prevalent methods for unified information extraction. a) pretraining and fine-tuing with SLM; b) instruction fine-tuing with LLM; c) inference with LLM and fine-tuning with SLM.

1 Introduction

Information Extraction (IE) endeavors to derive structured information from unstructured text (Andersen et al., 1992; Grishman, 2019), which involves a series of tasks, including named entity recognition, relation extraction, entity linking, aspect-based sentiment analysis, and event extraction (Muslea, 1999). Given its diverse objectives (entity, relation, event, etc.) and heterogeneous structures (spans, triplets, records, etc.), traditional IE methods often necessitate task-specific architectures and processes, entailing elaborate manual design (Grishman and Sundheim, 1996; Ji and Grishman, 2011). Despite some success, task-specific approaches impede rapid unified architectural development. Consequently, an alternative avenue of IE research focuses on addressing multiple subtasks using unified modeling architectures, as exemplified in recent works (Lu et al., 2022; Peng et al., 2023; Ping et al., 2023).

However, these prospective unified IE methods still grapple with several unresolved issues. One prominent challenge involves the noise interference introduced by negative samples during model training and prediction. Unlike traditional NLP tasks, there are usually long-tail data distributions in information extraction tasks that demonstrate imbalanced label quantities across various types, with a larger number of negative samples compared to positive ones (Huang et al., 2020; Dong et al., 2021; Liu et al., 2023). How to bridge label with output is also a challenge. Other than generative unified modeling architectures, Lin et al. (2020); Lou

Findings of the Association for Computational Linguistics: EACL 2024, pages 52–77
November 12-16, 2024

et al. (2023); Ping et al. (2023) employ extractive models to achieve unified information extraction through heterogeneous decoding processes across different subtasks. To capitalize on the knowledge acquired during the pretraining stage, many generative and extractive methods represent label types using natural language words. However, unlike context-based large language models such as GPT-3, PaLM, LLaMA, etc. (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023), the efficacy of smaller language models (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020) in comprehending abstract labels remains questioned. For instance, "Attack" is an event type hard to understand by a single word in ACE05-Evt, representing a range of conflict events such as wars, coups, strikes, terrorist attacks, etc., not merely its literal meaning.

Witnessing the remarkable performance of massively large language models in extensive NLP tasks, several LLM-based methods for information extraction have been proposed (Zhou et al., 2023; Wang et al., 2023b, 2022a; Wadhwa et al., 2023a; Gui et al., 2023; Wang et al., 2023c). However, there is still no optimal solution regarding the trade-off between effectiveness and efficiency, primarily due to the poor performance without fine-tuning in IE tasks (Han et al., 2023) and the overhead associated with training LLMs.

In this paper, we are dedicated to analysing these key problems and devising solutions. Through our investigation, we sum up three primary factors influencing the capabilities of unified IE models: 1) Noisy imbalanced data: a large number of negative samples and long-tail data distribution. 2) Abstract label type: obscure type words pose a challenge for understanding by LMs. 3) Diverse span granularity: annotated data from different sources has various criteria for identifying spans. Consequently, we posit that the primary capabilities of unified information extraction models revolve around anti-interference, label understanding, and span identification, addressing the aforementioned challenges. To tackle these issues, we propose the collaborative framework that consists of a Recognizer and multiple Experts. The Recognizer, an LLM proficient in anti-interference and label understanding, is tasked with recognizing label types and filtering negative samples. On the other hand, Expert utilize type indication as prompt to generate structured text, which are fine-tuned on low noise data distribution for a specific IE task. The Recognizer and Experts operate in a two-stage pipeline to produce general schemas for universal IE tasks, as illustrated in Figure 1. Different from previous research, our approach focuses more on solving the aforementioned problems and achieving performance improvement by simultaneously utilizing the potential advantages of LLM and SLM.

To validate the effectiveness and generality of GCIE, we conduct extensive experiments, encompassing 6 IE subtasks across various datasets. The experimental results demonstrate the rationality of key capabilities for unified IE and excellent performance under the supervised and few-shot settings. These findings collectively suggest that the integration of SLM and LLM yields enhanced information extraction capabilities.

In conclusion, the main contributions are summarized as follows:

1) We analyze the distinct benefits of context-based LLM and fine-tuned SLM for unified information extraction. We identify and articulate three essential capabilities that are crucial for addressing the fundamental challenges commonly encountered in universal IE tasks.

2) We propose the general collaborative framework for universal information extraction in a unified paradigm, designed to harness the complementary advantages of LLM and SLM to acquire the key capabilities.

3) We design task-specific prompts for negative samples filtering and type recognition of Recognizer and self-correction strategy for effective Expert learning.

4) We conduct a series of evaluation and exploration experiments to validate the rationality and effectiveness of our approach.

2 Key Capabilities for Unified Information Extraction

In this section, we outline the essential prerequisites for tackling the challenges inherent in universal information extraction tasks, delineating them into three key capabilities. We then elucidate the significance of these capabilities, underscoring why a robust IE model should incorporate all three. While our investigation is approached from a unified IE perspective, it is also applicable to numerous task-specific methodologies. The similar topic is discussed by (Sainz et al., 2023).

Anti-interference refers to the robustness of an

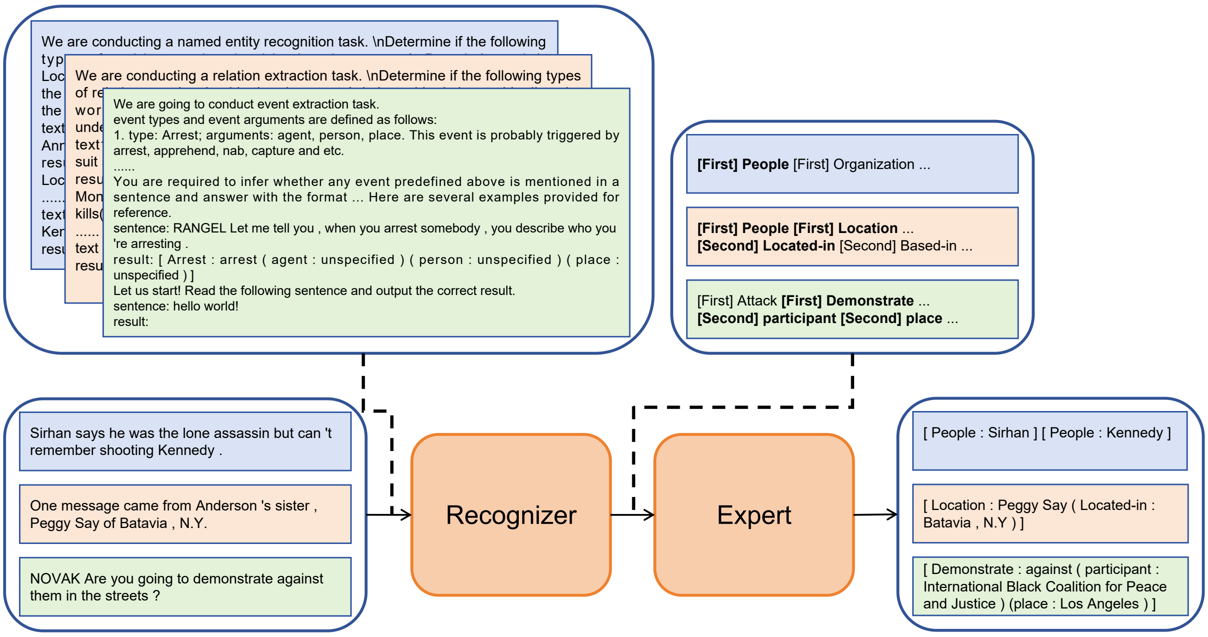

Figure 2: The overall architecture of GCIE that receives unstructured text and output task-specific schemas. In the prompts of Expert, types recognized by Recognizer are marked in bold. This framework can function in an end-to-end manner during the prediction phase.

IE model against noise in data distribution. In practice, many documents and sentences do not contain any predefined information element, often referred to as negative samples, which are considered noisy data. For instance, common IE datasets such as ACE2005 and SciERC contain a number of negative samples, which are relevant to event extraction, named entity recognition and relation extraction. To substantiate the significance of this capability, we perform the anti-interference test to evaluate the susceptibility of both LLM and SLM to negative samples (see Appendix B.1). Our findings indicate that the proportion of negative samples significantly influences the performance of generative IE models. Moreover, we can observe that SLM is more susceptible to interference than LLM.

Label-understanding describes the semantic understanding ability to predefined label. In recent years, many research works unlocked the label understanding ability of pretraining language model 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 primarily manifest in NLP tasks with simple label words such as 'positive', 'great', and 'person'. More abstract and polysemous label words are often too ambiguous for common language models to comprehend. In our exploration experiments (see Appendix D), we observe variations in model performance depending on the styles of label words replaced, ranging from simple capitalizations to other lexical alterations. This implies that SLM does not exhibit the same degree of sensitivity to abstract label words as LLM does.

Span-identification refers to the capability of accurately identifying information elements that likely represent entities, event triggers, or event arguments. To investigate this capability, we evaluate a context-based LLM and a fine-tuned SLM under different settings on the span identification task (see Appendix B.2). The performance of LLM in this regard is notably inferior to that of fine-tuned SLM. This discrepancy can be attributed to difference in dataset annotations, leading to diversity in span granularity. For instance, pairs such as "man" - "the man", "hospital in Boston" - "hospital", and "2 soldiers" - "soldiers" exemplify this variability. When subjected to rigorous evaluation metrics, LLM, lacking adaptation for a specific data distribution, struggles to match the performance of fine-tuned SLM.

3 General Collaborative Framework

In this section, we introduce a two-stage (Recognition & Filtering and Identification) general collaborative framework combining LLM and SLM to acquire capabilities 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, context collection and structure collection. Figure 6 demonstrates several examples for this unified modeling schema. Wherein the label collection includes predefined label type tokens, and the 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 type recognition and schema generation. Additionally, one point we consider very important, is the uniqueness of type words. For instance, we suggest type word "method" is substituted by "Methods", because the "method" in text typically is a entity with type of "Generic".

3.2 Framework Architecture

Our framework consists of Recognizer (black-box LLM only used for inference) and Expert (fine-tuned SLM), illustrated in Figure 2. In detail, Recognizer receives a sentence *s* and a task-specific instruction comprising examples *e* and the task question *q* as input. Utilizing a small set of input-output pairs for reference, the Recognizer generates the response to the question in the same format. The result given by Recognizer can be written as follows:

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

where $a = \{(typ_1, val_1), ..., (typ_n, val_n)\}$ is treated as a tuple collection with n type words and binary values, indicating which predefined types may exist in the sentence *s*.

In the designed task question *q*, each predefined label type is represented by a single word or a short phrase along with an interpretation. By associating these interpretations with examples, rather than relying solely on hard tokens as in SLM-only methods, LLM comprehends the actual semantics of each type more effectively, eliminating the concern of overfitting during model training. When choosing examples, it is crucial to consider different type combinations that enable LLM think comprehensively.

After the recognition process, with low confident types ($val_i = 0$) filtered out, high confident types ($val_i = 1$) organized as type indication (Expert prompt) are concatenated with sentence *s* as the input of Expert. We denote Expert prompt and sentence processed by tokenizer respectively as $p = \{p_1, p_2, ..., p_j\}$ and $t = \{t_1, t_2, ..., t_l\}$. The real words of Expert prompts used in our experiments for each dataset is listed in Appendix C. Theoretically, any auto-regression generative language model could be used as the base model of Expert, which predict conditional probability $\mathcal{P}(y_i|y_{<i}, p, t)$ of the next token y_i, given the context and input. Finally, when Expert finishes prediction when it generates the end symbol, we sample tokens by step from the logits to get the final output sequence o. The total generation process can be written as follows:

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

where $o = o_1, o_2, ..., o_k$ is the result of sampling with task-specific structured schema with sequence length *k*. *Decoder*$(\cdot)$ is the decoder of Expert and $o_i = Decoder(o_{<i}, p, t)$.

While generating structured schema rather than natural language text, a few sampling techniques could be applied when the Decoder of Expert operates, such as greedy search, beam search and constrained-decode (Lu et al., 2021). We try the three sampling techniques in our method, but no significant performance difference is observed. That is to say, our method does not depend on particular ways of decoding.

3.3 Expert Learning

To acquire span-identification capability, Expert requires a fine-tuning process. At present, we consider multiple feasible training plan which produces two bifurcation points. The first one is that whether Expert prompt from gold label or Recognizer prediction are used in training. The second one is that whether multiple task-specific Experts or a unified Expert for all IE tasks are maintained. We carry out thorough comparison about these issues in our supervised experiments. For simplicity, we assume $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ uniformly represent the train set of certain IE dataset. Therefore a most straightforward way to optimize parameters is minimizing the negative logarithmic

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)	92.44	76.90	86.24	93.35	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: The results of GCIE on NER, RETriplet and NER&RE tasks. We report the average F1 scores on 3 random seeds. †: The model has additional training process such as structure pretraining. ♣: The trainable model parameters (typically exceeding 10B) are an order of magnitude larger at least than that of ours. Task-specific IE models (upper part of the table) and unified IE models (lower part of the table) are separated with horizontal line.

likelihood expectation on train set:

$$\mathcal{L} = \sum_{(x,y)\in\mathcal{D}} -log\mathcal{P}(y|x,p;\theta) \tag{3}$$

where p is Expert prompt from Recognizer prediction or gold label and θ denotes all trainable parameters of Expert.

While training Expert using gold labels can reduce the expensive cost associated with LLM inference, it may lead to inconsistency between train data distribution and test data distribution. In view of this, unless specified otherwise, our training process use type indication from Recognizer prediction rather than gold label. Besides this, an inherent challenge in pipeline IE models is error propagation. Unlike inter-task pipeline models, GCIE operates as a general two-stage pipeline framework. The error propagation in GCIE can diminish its generalization ability due to its heavy reliance on type prompt derived from Recognizer prediction. Through Anti-interference test, we have drew the conclusion that fine-tuned SLM is more susceptible to indication omission than redundancy. To address this issue, we introduce the self-correction strategy to mitigate the Expert's over-dependency on type indication. Specifically, we introduce a reject probability subject to Bernoulli distribution, denoted by $P_r \sim Bernoulli(\alpha_r)$ for each predefined type across the all IE datasets. The value of α_r is determined by the recall score of Recognizer on development set. If certain type is not predicted by the Recognizer, it is excluded along with its reject probability from the Expert prompt. Under this self-correction mechanism, the initially deterministic type prompt becomes uncertain:

$$\mathcal{P}(p_i|x) = R_i + (1 - R_i) \cdot (1 - P_{ri}) \tag{4}$$

where $\mathcal{P}(\cdot)$ computes the conditional probability of p_i, which denotes the i-th type and x denotes the input sentence. $R_i \in \{0, 1\}$ is the prediction result of Recognizer of the i-th type.

In this way, the original Expert prompt p is replaced by $\tilde{p} = \{\tilde{p}_1, \tilde{p}_2, ..., \tilde{p}_{|p|}\}$, which is simultaneously robust to prediction errors and closer to the real results. Notably, $\mathcal{D}$ is replaced with $\tilde{\mathcal{D}}$ not containing any negative sample when self-correction mechanism is applied. Now the final optimization objective for Expert learning is:

$$\mathcal{L} = \sum_{(\tilde{x},\tilde{y})\in\tilde{\mathcal{D}}} -log\mathcal{P}(\tilde{y}|\tilde{x},\tilde{p};\theta) \tag{5}$$

4 Experiments

To validate the efficacy of the proposed methodology and explore pivotal factors within the GCIE framework, we systematically conduct an extensive series of experiments. These experiments encompassed the performance evaluation of GCIE and the exploratory investigations regarding Recognizer,

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	56.20	-	-	-	-	-	-
(Wang et al., 2023b)♣	-	77.13	**72.94**	67.80	63.53	-	-	-	-
GCIE w/o SC (ours)	81.13	81.68	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: The results of GCIE on ED, EE and ABSA tasks. We report the average F1 scores on 3 random seeds. †: The model has additional training process such as structure pretraining. ♣: The trainable model parameters (typically exceeding 10B) are an order of magnitude larger at least than that of ours. Task-specific IE models (upper part of the table) and unified IE models (lower part of the table) are separated with horizontal line.

and Expert. In all experiments, the default base model for Expert is Flan-T5 (Shen et al., 2023a), while LLM refers to Claude2.1.1 [1]. The detail experimental configuration can be found in the Appendix C.

4.1 Experimental Settings

Task. We select 6 representative IE tasks: 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). The comprehensive performance evaluation of GCIE and its variants (without filtering, self-correction and unifying) is carried out. Moreover, a few designed tasks including negative samples recognition, type recognition and span identification are involved.

Datasets. In our experiments, all datasets used in the supervised, few-shot settings and exploration experiments include 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), CommodityNews (Lee et al., 2021), SemEval-14 (Pontiki et al., 2014), SemEval-15 (Pontiki et al., 2015), SemEval-16 (Pontiki et al., 2016). For the aforementioned tasks and datasets, the detailed statistical information is described in Appendix A.

4.2 Experiments on GCIE

4.2.1 Supervised Settings

The main results from the performance evaluation of GCIE on supervised settings are shown as Table 1 and Table 2. Specifically, GCIE variants, such as GCIE-unify, denoting the unified model across all datasets, SC, representing the self-correction strategy, and F, representing the negative sample filtering mechanism, are examined. GCIE achieves quite impressive scores across 6 IE tasks. For most of these datasets, our method surpasses all unified IE methods including SLM-only and LLM-only models. Especially on partial datasets, such as CoNLL03 (NER), ACE05-Evt (ED), CASIE (EE) and 14lap/16res (ABSA), GCIE achieves state-of-the-art performance. Only on a few datasets, our method slightly underperforms baselines. Additionally, we try to maintain a unified set of parameters for all IE tasks (GCIE-unify). In this case, we observe a slight decrease in model performance across all datasets, but it still remains 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 comparable to, even exceeding state of the art IE models

[1] https://claude.ai/

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: The results of GCIE and baselines on few-shot settings.

with fewer training parameters, which benefits from collaboration of LLM and SLM in negative sample filtering, type recognition and self-correction strategy.
(2) Compared to baselines, the improvement on performance of our method varies significantly across different tasks and datasets. For example, GCIE outperforms task-specific models and unified models on ACE05-Evt (ED) and CoNLL03 (NER), but it struggles to compete with SOTA model on CoNLL04, SciERC and ACE05-Evt. We attribute this phenomenon to three main reasons: dataset preference, capacity range of our method and prompt design. We discuss detailedly these factors in Error Analysis (Section 5) and Reason Analysis (Appendix D.3).
(3) All modules including Recognizer(recognition and filtering), Expert and self-correction strategy of our framework play important roles. Specially, self-correction mechanism is capable of correcting the reliance of Expert on type indication, and omitting it would result in a huge performance drop.
(4) We try to train a unified Expert for all tasks and datasets and find a little performance decline. We speculate that it is due to the lack of uniformity in type definition and span granularity over different datasets.

4.2.2 Few-shot Settings

To explore the performance of GCIE in resource-constrained scenarios, we randomly sample from the train set in both 1-shot and 10-shot settings for each IE task, and evaluate on full-sample test set. We repeat each experiment 10 times and employ the same evaluation metrics used in supervised settings. Without type indication from Recognizer, the Expert instead utilizes SSI and SEL, as proposed by UIE (Lu et al., 2022). Flan-T5 operates with fixed type indication. As depicted in Table 3, GCIE demonstrates significant outperformance compared to both Flan-T5 and Expert across all datasets. We observe that, particularly in complex structured tasks such as event extraction, both Flan-T5 and Expert struggle to effectively learn the correct input-to-output dependency in the absence of type indication, rendering them vulnerable to overfitting. In contrast, Recognizer enhances the robustness of GCIE through only a few demonstrations to identify potential types and negative samples.

4.3 Experiments on Recognizer

The overall performance of GCIE is significantly contingent upon 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. This task aims to ascertain the presence of predefined types within a given text. We conceptualize type recognition as a multi-label classification task and adopt the F1 score as the primary evaluation metric.

Due to the variances in structures and objectives across different Information Extraction (IE) subtasks, we craft distinct instructions for Claude2.1 prompts tailored to each IE subtask (For detailed information, refer to Appendix C). Each instruction includes a task-specific question and several examples, serving as hyperparameters in the Recognizer module. Additionally, we fine-tune two SLM methods as baselines for comparison. To validate generality, We also explore this ability on other LLMs in Appendix D.4.

Considering the influence of input length on LLM performance, we set maximum values for the number of demonstration for each dataset. In Table 4, it is evident that as the number of examples increases, Claude2.1 consistently exhibits an upward trend in performance. And with the increasing number of examples, Claude2.1 demonstrates notable performance improvements compared to fine-tuned SLMs across almost all IE subtasks, particularly in challenging tasks such as event extraction. Notably, Claude2.1 exhibits significantly higher recall scores than precision across all datasets, suggesting that LLM recognizes types with a high level of confidence. In summary, we can draw conclusions as follows:
(1) Claude2.1 outperforms fine-tuned SLM by a large margin, especially in complex tasks, due to its superior label-understanding and anti-interference abilities. In addition, during the experiment process, we observe that Claude2.1 makes the prediction with high confidence and some inference steps.
(2) LLM serves as the type recognizer, achieving

Dataset	Element	n	Roberta-large			Generation-based			Claude2.1 k=5			Claude2.1 k=10		
			P	R	F	P	R	F	P	R	F	P	R	F
CoNLL03	Ent	30	92.6	90.8	91.7	92.6	94.4	93.5	91.3	96.3	93.7	93.4	98.6	**95.9**
SciERC	Ent	30	70.2	63.3	66.6	65.9	79.2	71.9	71.4	83.3	**76.9**	-	-	-
ACE05-Rel	Ent	40	88.6	84.8	86.7	84.7	92.2	88.3	78.6	94.2	85.7	82.6	96.4	**89.0**
CoNLL04	Ent	40	84.7	87.1	85.9	92.4	95.9	94.2	90.6	98.0	94.2	93.4	98.0	**95.6**
	Rel	30	79.4	77.0	78.2	87.9	84.4	**86.2**	80.0	90.6	85.0	-	-	-
ACE05-Evt	Evt	100	86.7	82.3	84.4	86.6	66.3	75.1	88.1	96.4	**92.1**	-	-	-
	Arg	80	69.0	63.3	66.0	71.0	50.1	58.3	73.3	83.3	**78.0**	-	-	-
14-lap	Sen	30	89.2	83.7	86.4	70.5	70.3	70.4	79.5	96.3	87.1	84.1	98.1	**90.6**

Table 4: The results of type recognition of Roberta and Claude2.1 on the dev sets of various datasets. Roberta-large is designed to directly model types while the Generation-based method represents the structured generation paradigm used by some task-specific and unified IE methods. For the Generation-based method, we only take types into account to calculate the scores. These two SLMs are fine-tuned on full-sample train set for each datatset. n is the maximum value of example number.

remarkable results across many datasets with only a limited number of examples. In practice, it is worth considering leveraging the high recall property of LLM to guide SLM extraction.

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

5 Error Analysis

In this section, we discuss the reason why our method can achieve remarkable performance improvements through quantitative error analysis. We report the results of GCIE and the baseline on several types shown in Table 5. The baseline which is the Expert-only model as same as that we use in the few-shot setting underperforms GCIE on all of these types. But the improvements vary dramatically among different types. Specifically, GCIE exceeds the baseline by a large margin on event triggers and arguments. These elements are either polysemous or low-resource, which makes it difficult to be captured by Expert. Instead, Recognizer can understand complex meanings based on fewer supervised data and indicate Expert to identify them. GCIE also leads the baseline on common sentiment, entities and relations by a moderate margin. These types include sufficient supervised data, which mitigates the impact of not having Recognizer, but their performances still can be improved by using LLM. However, there are other special types of performance where GCIE and the baseline are very close, such as "Generic" and "Cell type". GCIE get less benefits from Recognizer due to LLM not specializing in some professional fields and recognizing unspecific types. We summarize the main points of error analysis as follows:

1) GCIE benefits much from the collaboration of Recognizer and Expert, overwhelming the Expert-only baseline on many types.
2) Recognizer helps to significantly improve the recognition on polysemous or low-resource types.
3) The collaborative framework also can further improve the performance on common types although they have sufficient supervised data.
4) Recognizer is not good at recognizing general types such as "Generic", "Miscellaneous" and etc. and performs poorly in some professional fields such as "biology", "science", etc..

Dataset	Element	Type Name	Baseline	GCIE
GENIA	Ent	Protein	77.46	**83.19**
	Ent	Cell type	76.42	**78.76**
SciERC	Ent	Generic	64.98	**65.17**
CoNLL04	Rel	Live in	75.76	**77.16**
	Rel	Located in	74.99	**78.16**
	Rel	Work for	70.13	**76.51**
NYT	Rtp	Place of birth	93.07	**94.45**
15res	AO	Aspect-Opinion	65.60	**70.72**
	Sen	Positive	70.39	**76.12**
ACE05-Evt	Evt	Justice.Trial-Heraring	50.0	**100.0**
	Evt	Justice.Acquit	0.0	**100.0**
	Evt	Meet.Contact	70.83	**80.80**
	Evt	Conflict.Attack	73.12	**88.76**
	Arg	Defendant	50.0	**69.26**
	Arg	Destination	46.34	**66.67**

Table 5: The F1 scores of GCIE and the baseline reported on several selected types.

6 Related Work

From the perspective of the target tasks, we primarily present research works about various paradigms for information extraction. 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 deep language models and the increasing demand of heterogeneous information processing, more and more IE models are designed in the unified paradigm to address various IE tasks. Early unified IE models typically employ multi-task joint training to enable the model to adapt various information extraction tasks with different objectives and schemas (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. Some recent research efforts (Peng et al., 2023; Ping et al., 2023; Gao et al., 2023) aim to introduce novel methods to adapt universal IE tasks rather than unified modeling. However, the most closely related approaches to our work are the unified structured generation paradigm for a range of IE tasks (Lu et al., 2022; Wang et al., 2022a, 2023b). Since the advent of ChatGPT and other LLMs, more and more researchers take efforts to unlock the potential of LLMs and bridge the performance gap with SOTA results in IE tasks (Gui et al., 2023; Wang et al., 2023c; Wadhwa et al., 2023b; Sainz et al., 2023; Shen et al., 2023b). We also regard this as a prospective research direction of unified information extraction.

7 Conclusion

In this study, we analyze the important factors for information extraction and introduce three core capabilities. These capabilities, typically not concurrently possessed by existing IE models, are identified through a series of exploration experiments. Our findings suggest that context-based LLM is proficient in identifying negative samples and recognizing predefined types. Building upon this insight, we propose GCIE for unified information extraction, which combines the strengths of LLM and Experts to encompass both of these capabilities. Extensive experiments validate that, compared to existing LLM-only and SLM-only methods, GCIE exhibits excellent performance across many IE tasks. All of these indicate a prospective unified IE research direction to take advantages of LLM and fine-tuned SLM.

Limitations

Despite the success of our approach, some limitations should be pointed out and addressed in the future:

1) Our approach requires some additional inference latency brought by LLM compared to SLM-only methods.

2) Designed prompt is one of the important factors that influence the performance and stability of Recognizer.

3) The hyperparameters in self-correction mechanism are determined manually, which is likely to be sub-optimal.

4) Our method has the property of dataset preference, which makes it perform mediocre on certain datasets.

5) We haven't explore more extensive scenarios, such as open information extraction tasks.

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.

T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Askell Amanda, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Henighan Tom, Rewon Child, A. Ramesh, DanielM. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, EricJ. Sigler, Mateusz Litwin, Scott Gray, Chess Benjamin, Jack Clark, Christopher Berner, McCandlish Sam, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv: Computation and Language,arXiv: Computation and Language*.

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.

Meisin Lee, Lay-Ki Soon, Eu-Gene Siew, and LyFie Sugianto. 2021. An annotated commodity news corpus for event extraction. *arXiv: Computation and Language,arXiv: Computation and Language*.

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.

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, Harris 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.

Oscar Sainz, Iker Garcia-Ferrero, Rodrigo Agerri, Oier-Lopezde Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction.

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. 2023a. 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, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023b. Promptner: Prompt locating and typing for named entity recognition.

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, our experimental resources involves several datasets across 6 information extraction tasks. We provide the detailed description of each task, dataset, and evaluation metric as follows. The detail statistics of all IE datasets used in our experiments are listed in Table 6.

Named Entity Recognition is a task in NLP 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 NLP 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.

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.

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
CommodityNews	19 Evt	1245	-	311
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 6: The statistics of all IE datasets used in this study.

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 Capability Test

In this section, we discuss the three key abilities through quantitative experiments and make a comparison between LLM and SLM. Because it is hard to directly compare the performance of LLM and SLM in the aspect of Label-understanding, we use a ablation experiment (see D) to prove the conclusion that SLM is not as sensitive to the label style as context-based LLM in the process of fine-tuning.

B.1 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,

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	Claude2.1	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 7: Hyper-parameters for GCIE training on both supervised and few-shot settings.

we fine-tune small language model with structural 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 fine-tuned models 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 9, 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.

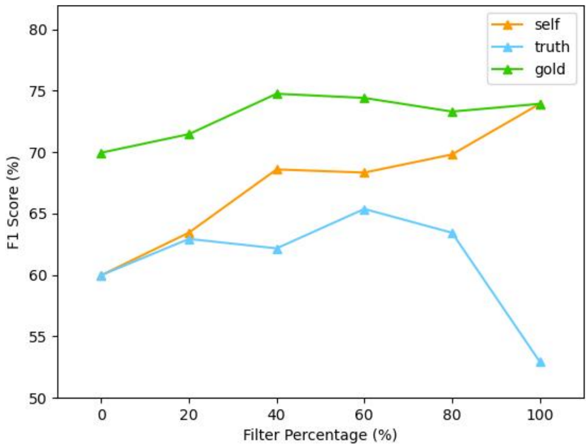

Figure 3: The performance on ACE05-Evt of generative fine-tuned models with negative sample filteration in varying proportions. "self" denotes the score on simulated labels by random sampling at the ratio; "truth" denotes the score on practical labels; "gold" denotes the score on positive labels.

B.2 Span-identification Test

To compare this ability between LLM and SLM, we design the span identification task based on 3 datasets across 3 information elements. In specific, this task ask models to generate the true spans (we select three common information elements: event trigger, entity and opinion) given the type indicators (we indicate LLM with more informative prompts than SLM). We choose GPT-3.5-turbo as LLM and fine-tune Flan-T5 as SLM. It can be seen that, under all settings the F1 scores of fine-tuned T5 outperform that of GPT-3.5-turbo by a large margin. Although increasing the number of few-shot can improve performance, it is also clear that only augmenting the context does not make LLM compete with fine-tuned SLM on span identification task.

C Experiment Details

In this section, we describe details of experiments that include hyper-parameters on supervised and

Model	CommodityNews Trigger	CoNLL04 Entity	15-res Opinion
GPT-3.5-turbo k=2	80.85	75.93	62.87
GPT-3.5-turbo k=5	83.71	77.55	64.86
T5-base	92.12	87.59	75.21
T5-large	96.09	90.45	78.29

Table 8: The results of span identification based on ChatGPT and fine-tuned T5. k is the number of few-shot.

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
Claude2.1 k=5	85.5	75.0	95.0

Table 9: 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.

few-shot settings, Recognizer prompt construction and Expert prompt construction.

C.1 Hyper-parameters

As shown in Table 7, on supervised and few-shot experiments, we select Claude2.1 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, we suggest a 48G memory is accessible at least.

C.2 Recognizer Prompt

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

CoNLL04
We are conducting joint entity and relation extraction task.
entity types are defined as follows:
People(people), Location(location), Organization(organization), Others(other entity such as time).
relation types are defined as follows:
Based-in(organization is based in location), Located-in(location is located in location), Live-in(people lives in location), Work-for(people works for organization), Kill(people kills people).
Please determine if there exist entities and relations defined above in the given text. Referring to several following examples, complete the content of result.
text:" U.S. decision-makers should understand that the signals they send today will have major ramifications for the Israeli approach to the Arrow program , " says Marvin Feuerwerger in a 1991 study for the Washington Institute for Near East Policy .
result:People(Marvin Feuerwerger), Location(U.S.), Organization(Washington Institute for Near East Policy), Others(1991); Based-in(absence), Located-in(absence), Live-in(absence), Work-for(Marvin Feuerwerger works for Washington Institute for Near East Policy), Kill(absence).
text:Meanwhile , on a separate occasion , Prince Ranariddh , first prime of Cambodia , reiterated the Phnom Penh government 's wish to open a Cambodian Embassy in Jakarta as soon as possible .
result:People(Prince Ranariddh), Location(Cambodia)(Jakarta), Organization(Phnom Penh government), Others(absence); Based-in(Phnom Penh government is based in Cambodia), Located-in(absence), Live-in(Prince Ranariddh live in Cambodia), Work-for(Prince Ranariddh works for Phnom Penh government), Kill(absence).
text:He graduated from high school from Benton , Tenn. and from Tennessee Tech in Cookville , and holds a doctorate in physics from Virginia Tech .
result:People(absence), Location(Benton)(Tenn. Cookville), Organization(Tennessee Tech)(Virginia

Tech), Others(absence); Based-in(Tennessee Tech is based in Tenn.)(Tennessee Tech is based in Cookville), Located-in(Benton is located in Tenn.)(Cookville is located in Tenn.), Live-in(absence), Work-for(absence), Kill(absence).
text:In 1752 , flagmaker Betsy Ross was born in Philadelphia .
result:People(Betsy Ross), Location(Philadelphia), Organization(absence), Others(absence); Based-in(absence), Located-in(absence), Live-in(Betsy Ross lives in Philadelphia), Work-for(absence), Kill(absence).
......
Let us start! Please analyse the following sentence and complete the result.
text:adviser to PLO Chairman Yasir 'Arafat by Sa 'id Mu 'addi in Cairo on 18 May from the " With the Midday Events " program – recorded) (Excerpt) (passage omitted) (Mu 'addi) One last question , Dr. Nabil .
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: time, tool, attacker, victim, purpose, place, damage amount, number of victim, number of data
2. Phishing: place, purpose, damage amount, trusted entity, attack pattern, attacker, victim, time
3. Ransom: victim, attacker, place, time, attack pattern, payment method, some financial and person data, tool, damage amount
4. Discover Vulnerability: vulnerability, vulnerable system owner, vulnerable system, time, common vulnerabilities and exposures, supported platform, vulnerable system version, capabilities
5. Patch Vulnerability: time, vulnerable system version, common vulnerabilities and exposures, patch, patch number, releaser, The open source content management project, supported platform, vulnerability, vulnerable system, issues addressed
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)].
Here are several examples.
demonstration 1
sentence: As of Saturday , Atlanta officials and federal partners were still “ working around the clock ” to resolve the ransomware attack on city computers that occurred around 5 a.m. on Thursday , March 22 , and encrypted some financial and person data .
result: [ransom : the ransomware attack (victim : city computers) (time : 5 a.m. on Thursday , March 22) (attack pattern : encrypted some financial and person data)] [discover vulnerability : none] [data breach : none] [patch vulnerability : none] [phishing : none]
demonstration 2
sentence: The open source content management project has issued an unscheduled security update to augment its previous patch for Drupalgeddon2 .
result: [data breach : none] [ransom : none] [patch vulnerability : has issued (patch : its previous patch) (releaser : The open source content management project) (vulnerable system : Drupalgeddon2)] [discover vulnerability : none] [phishing : none]
demonstration 3
sentence: Bleeping Computer , too , has spotted increases in phishing campaigns targeting Blockchain.info in December 2016 and December 2017 .
result: [data breach : none] [discover vulnerability : none] [ransom : none] [phishing : phishing campaigns (trusted entity : Blockchain.info) (time : December 2016 and December 2017)] [patch vulnerability : none]
demonstration 4
sentence: Google also provided Microsoft with an additional 14 - day grace period to have a fix available for its monthly Patch Tuesday release in February , but Microsoft missed this goal because “ the fix is more complex than initially anticipated . ”
result: [ransom : none] [discover vulnerability : none] [data breach : none] [phishing : none] [patch vulnerability : available (patch : release) (releaser : Microsoft) (time : February)]
......
Let us start! Read the sentence and complete content of the result. You should think step by step.
sentence: Ticketfly did n't comment on whether

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: The examples about overfitting of a fine-tuned generative IE model on negative samples .

any user information , such as credit card data , had been stolen in the cyberattack .
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, other.

CoNLL04: Person, Organization, Location, Other, 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,

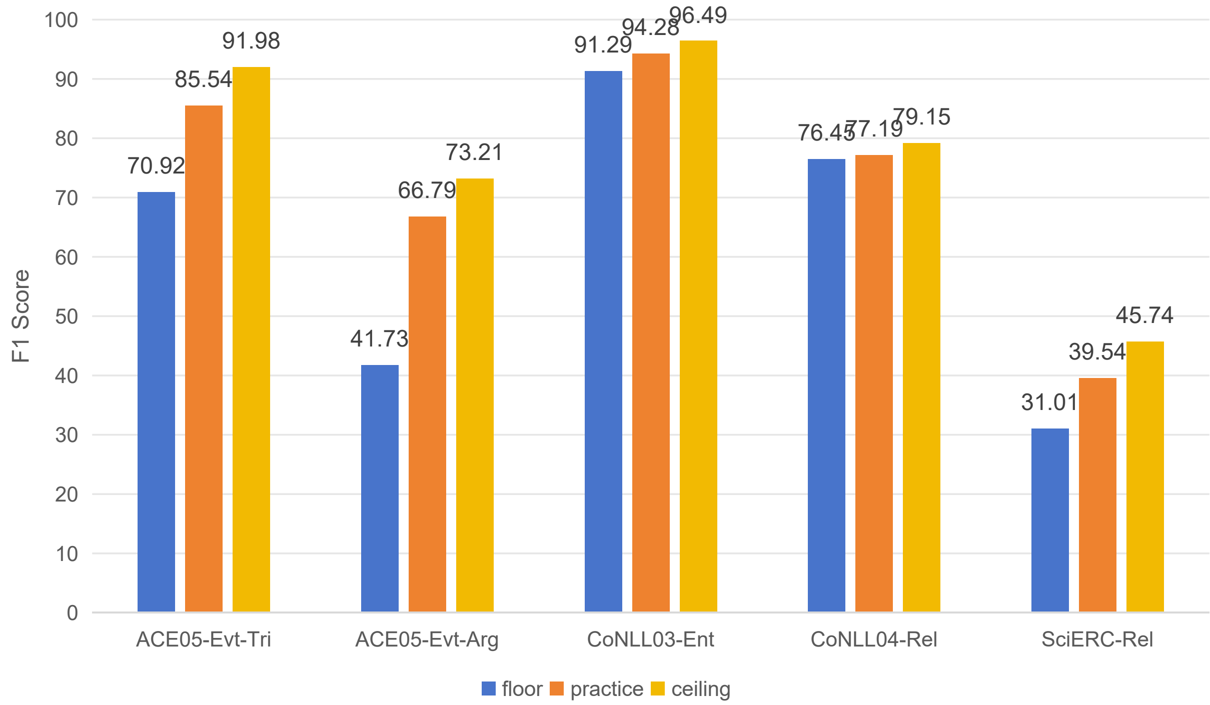

Figure 5: The capacity range of our framework based on Flan-T5-large. The "floor" denotes the minimum (fixed prompt), the "practice" denotes the practical results, and the "ceiling" denotes the maximum (optimal type indication and filtering).

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.

Model	**CoNLL03** Ent	**SciERC** Rel	**14-res** Sen
Claude2.1 w/o Span	89.8	57.2	86.1
Claude2.1	93.7	65.7	87.0
△Gain	+3.9	+8.5	+0.9

Table 10: The experimental results of the example format on entity, relation and sentiment.

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 10. 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 Claude2.1 needs more contextual information to understand the label semantics. Leveraging span mentions reasonably enhances the in-context learning ability of LLM, analogous to

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 11: The experimental results of the Type Phrase on CoNLL04 and SciERC dev sets.

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 of which are shown in Table 11. 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.

D.3 Reason Analysis

In this section, we analyze the factors leading to performance difference on different tasks and datasets. In terms of the results in supervised settings, our method performs excellently on event extraction and named entity recognition, but relatively poor on joint entity and relation extraction. The overall performance of GCIE depends on Recognizer and Expert, hence we examine them respectively and summarize the main reasons as follows:

Dataset Preference. Regarding Recognizer, we examine the results and failure cases on type recognition task. We find that LLM performs worse in analyzing relations between entities as expected. On CoNLL04, Claude2.1 used as our Recognizer sometimes overreason potential relations. For instance, given a sentence: *"On this date : In 1833 , Benjamin Harrison , the 23rd President of the United States , was born in North Bend , Ohio ."*, there is one relation *"(Benjamin Harrison; live in; North Bend , Ohio)"* annotated in the gold label. Other than this, Claude2.1 also predicts another unannotated relation *"(North Bend , Ohio; located in; the United States)"*, which is although known by us. As opposed to the gold label, overreasoning brings some false type indication leading to performance decline. On SciERC, there are generic entity types (Generic and OtherScientificTerm). We observe that Claude2.1 has weaker ability to recognize the relation when the head entity or the tail entity belongs to generic types. We named this phenomenon dataset preference since these are primarily decided by the intrinsic properties of dataset and LLM itself. As for ACE05-Evt, on the one hand, the dataset suffers a more severe long-tail distribution issue than others. Many event arguments only occur very few times throughout the training set (approximately one-fourth of the event arguments occur no more than five times). This unbalanced data distribution influences not only the quality of type recognition but also the quality of span identification. On the other hand, although LLM is able to understand labels to some extent, we find it still performs badly on some types. Similar with SciERC, the type "Agent" in event "Movement" is hard to be correctly predicted by LLMs including Claude, GPT-3.5 and Gemini-pro, because they seem to be unaware of what the argument represents.

Capacity Range. We posit that the performance achievable by our method is constrained within the measurable capacity range of our modeling architecture. By manipulating various components within our framework, we can ascertain the theoretical upper and lower bounds of our method's efficacy. Specifically, in scenarios where the Expert is constant, we postulate that the optimal type indication is derived from gold label, whereas fixed prompt serve as a baseline. Additionally, we exclude all negative samples in the optimal configuration. Figure 5 shows the capacity ranges of GCIE on 4 datasets. It's obvious that the improvement is constrained by the theoretical maximum and increases with the capacity range on certain dataset. This finding also explains why performance improvement of GCIE appears diverse over different datasets.

Prompt Design. It is universally acknowledged that different prompt of LLM leads to significant performance difference. In terms of our method, the task-specific question that encompasses type de-

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

scription and task instruction is the primary factor while the set of demonstrations remains unchanged. Although more than one questions are observed to be able to prompt LLM well, we also find different results between them. Especially, LLM is more sensitive to the type description for relation than that for entity. Task instruction for event extraction is more important than others. Even though prompt design is not as straightforward as other factors to influence the performance, we take it into consideration in view of the property of LLM.

D.4 Generality of Type Recognition Ability

In this part, we explore the generality of type recognition ability on some prevalent LLMs other than Claude. In specific, we choose GPT-3.5-turbo [1], Gemini-pro [1] and Mixtral-MoE [1] as test objects to perform type recognition on different IE datasets. For efficient evaluation, we randomly select 650 samples as test sample collection from the development set of each dataset. From the results of Table 12, GPT-3.5-turbo and Gemini-pro achieve the similar performance with Claude on both CASIE and ACE05-Rel, while Mixtral get the poor performance. We observe that Mixtral seems to suffer serious hallucination and does not output the correct schemas format demonstrated by few-shot examples on a lot of test samples of CASIE. This implies Mixtral is not able to perform type recognition on event extraction task. In addition, recall is more important than other metrics to our approach.

LLM	**CASIE-Arg**			**ACE05-Rel-Ent**		
	P	R	F	P	R	F
GPT-3.5-turbo	63.96	68.26	66.04	79.89	88.89	84.15
Gemini-pro	61.16	70.62	65.55	76.67	94.90	84.82
Mixtral-8*7B	-	-	-	78.48	64.58	70.85
Claude (reference)	67.88	73.71	70.67	78.62	94.21	85.71

Table 12: The comparison results of type recognition tasks on different LLMs. All LLMs use the same task instructions and are based on 5-shot context setting.

E Schema Format

The output formats utilized by both Recognizer and Expert adhere to the structures depicted in Figure 6. It is important to highlight that the outputs generated by LLMs needn't to provide accurate structured answers. Instead, it aims to discern the relevant information outlined within a given sentence. But presenting a comprehensive response which incorporates complete schemas as interpretable evidences, can facilitate LLMs to think step by step.

[1]https://openai.com/chatgpt

[1]https://gemini.google.com/

[1]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1