SCIR: A Self-Correcting Iterative Refinement Framework for Enhanced Information Extraction Based on Schema

Yushen Fang, Jianjun Li*, Mingqian Ding, Chang Liu, Xinchi Zou, Wenqi Yang

School of Computer Science and Technology, Huazhong University of Science and Technology
{yushen_fang, jianjunli, mingqianding, diu, xinchizou, yangwenqi}@hust.edu.cn

Abstract

Although Large language Model (LLM)-powered information extraction (IE) systems have shown impressive capabilities, current fine-tuning paradigms face two major limitations: high training costs and difficulties in aligning with LLM preferences. To address these issues, we propose a novel universal IE paradigm—the Self-Correcting Iterative Refinement (SCIR) framework—along with a Multi-task Bilingual (Chinese-English) Self-Correcting (MBSC) dataset containing over 100,000 entries. The SCIR framework achieves plug-and-play compatibility with existing LLMs and IE systems through its Dual-Path Self-Correcting module and feedback-driven optimization, thereby significantly reducing training costs. Concurrently, the MBSC dataset tackles the challenge of preference alignment by indirectly distilling GPT-4's capabilities into IE result detection models. Experimental results demonstrate that SCIR outperforms state-of-the-art IE methods across three key tasks— named entity recognition, relation extraction, and event extraction—achieving a 5.27 percent average improvement in span-based Micro-F1 while reducing training costs by 87 percent compared to baseline approaches. These advancements not only enhance the flexibility and accuracy of IE systems but also pave the way for lightweight and efficient IE paradigms.

Code & Datasets & Extended version — https://github.com/Franklin-Fang/SCIR

Introduction

Information Extraction (IE) stands as a pivotal technology within the field of Natural Language Processing (NLP), dedicated to automatically extracting and structure key information from unstructured text (Wilks 1997). Its research primarily revolves around three fundamental tasks: named entity recognition (NER), relation extraction (RE), and event extraction (EE). In recent years, as LLMs increasingly establish themselves as mainstream solutions for NLP tasks (Gu and Dao 2023; Jiang et al. 2025; Zhang et al. 2023; Xiong et al. 2025; Wan et al. 2024; Yang, Tang, and Tam 2023; Han et al. 2025; Mao et al. 2025), there has been a burgeoning academic interest in exploring the full potential and boundaries of LLMs within the IE domain. Several notable advancements have emerged in this pursuit (Xiao et al. 2023; Gui et al. 2024; Luo et al. 2024; Li et al. 2025). For instance, OneKE (Luo et al. 2024) achieved significant performance enhancements by leveraging data synthesis and fine-tuning strategies tailored to IE tasks. ChunkUIE (Li et al. 2025) introduced an innovative chunk-based extraction methodology, offering a fresh perspective on structured information retrieval from text. Furthermore, RUIE (Liao et al. 2025) integrated retrieval augmentation techniques into IE workflows, setting a new state-of-the-art benchmark.

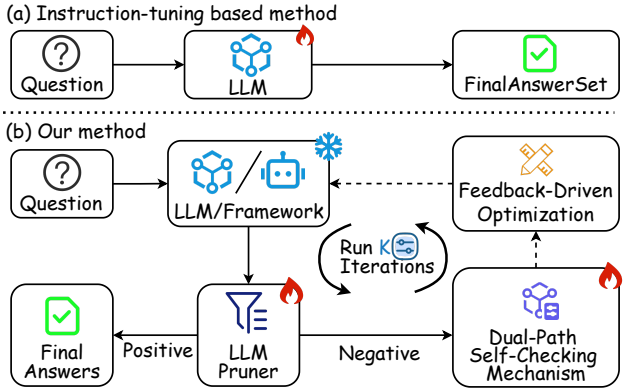

Figure 1: Traditional IE methods vs. our SCIR framework.

Despite their promise, LLM-based IE models still face two major challenges: (1) **High training costs and limited model flexibility.** Current mainstream methods predominantly rely on fine-tuning techniques to enhance domain-specific performance. However, this approach not only demands substantial computational resources and time, but also weakens the model's semantic understanding capabilities and restricts its generalization performance in new domains. More critically, existing frameworks are often tightly coupled with specific underlying models, making it difficult to adapt to their rapid iteration cycles (e.g., the GPT series updates every 3-6 months). The high cost of retraining, often taking weeks or even months, further impedes users from promptly adopting newer and more advanced models. (2) **Difficulty in aligning model preferences**. Existing information extraction models are significantly constrained by the inherent biases and blind spots present in human anno-

*Corresponding author.

tations. For edge cases that are often overlooked or prone to errors during the annotation process, these models lack remedial mechanisms and cannot completely eliminate such errors simply by increasing the volume of data. Furthermore, traditional supervised training adheres to a "static annotation to static inference" paradigm, lacking dynamic feedback and self-correction capabilities for addressing model error patterns. This hinders the models' ability to effectively enhance the accuracy and consistency of their outputs when confronted with unknown or complex contexts.

To address these challenges, we propose a novel IE paradigm (as depicted in Figure 1) and introduce the Self-Correcting Iterative Refinement (SCIR) framework, which achieves breakthroughs through the following innovations. Specifically, to tackle the first challenge, we designed the SCIR framework to eliminate the need for fine-tuning extraction models. This framework leverages a Dual-Path Self-Correcting mechanism and a Feedback-Driven Optimization mechanism to enhance the flexibility of IE systems. The Dual-Path Self-Correcting mechanism verifies the completeness of extraction results through two pathways: redundancy detection and missing detection. The Feedback-Driven Optimization mechanism generates iterative prompts based on verification results to drive a context-learning-based iterative generation process. This design enables flexible substitution of IE models while requiring only a single training session for the Dual-Path Self-Correcting mechanism, regardless of model replacements—significantly enhancing system flexibility. To address the second challenge, we constructed a Multi-task Bilingual Self-Correcting (MBSC) training set based on the IEPile dataset (Gui et al. 2024), specifically designed for model preference alignment training. Unlike traditional static datasets reliant on manual annotations, the MBSC dataset centers on error instances generated by GPT-4 in information extraction tasks. It systematically collects edge cases often overlooked, annotation blind spots, and model error-prone points, followed by multi-task labeling. By incorporating real-world error scenarios, MBSC enhances the diversity of training samples, enabling models trained on MBSC to identify biases in extraction results and provide dynamic feedback signals to extraction models. Our main contributions can be summarized as follows:

- **Framework Paradigm Shift**: We propose SCIR, a pioneering fine-tuning-free IE paradigm that achieves exceptional generalization via integrating Dual-Path Self-Correcting and Feedback-Driven Optimization mechanisms, enabling seamless IE base model substitution and iterative refinement while ensuring cost efficiency.
- **Specialized Dataset Synthesis**: We introduce MBSC, an innovative dataset tailored for error correction and preference alignment in IE models, systematically capturing edge cases, annotation blind spots, and model errors to enhance training diversity and robustness.
- **Empirical Performance Breakthrough**: Through comprehensive zero-shot transfer evaluations across 11 multilingual benchmarks, we demonstrate SCIR's outstanding effectiveness with an 5.27% average F1-score increase, underscoring its potential to revolutionize IE by providing a plug-and-play solution that cuts training costs while maintaining high performance.

Related Work

The evolution of information extraction (IE) has undergone three distinct phases. Early rule-based systems (Chiticariu, Li, and Reiss 2013; Maturana, Riveros, and Vrgoč 2017; Thenmozhi and Aravindan 2018) utilized manually engineered patterns (e.g., regular expressions) for domain-specific tasks, but their poor generalization across domains and high maintenance costs motivated the shift toward statistical and machine learning approaches. Methods such as Hidden Markov Models and Support Vector Machines (Cortes and Vapnik 1995) leveraged annotated corpora to improve model adaptability. Currently, deep learning dominates IE research, with Transformer-based architectures like BERT (Devlin et al. 2019) achieving state-of-the-art performance through large-scale pretraining, particularly in relation and event extraction tasks.

Modern IE frameworks primarily follow two paradigms: open IE and schema-based IE (Qi et al. 2024). Open IE systems extract unstructured semantic triples (e.g., for question answering) without predefined schemas (Lou et al. 2023), necessitating post-hoc standardization through clustering or alignment techniques (Ma et al. 2023; Lu et al. 2022). In contrast, schema-based IE—common in specialized domains—adheres to structured templates (e.g., entity-relation taxonomies or event hierarchies) to ensure extraction precision and interoperability (Tedeschi and Navigli 2022; Zheng et al. 2017). These schemas range from flat entity lists to complex nested structures, providing explicit constraints to guide the extraction process (Wang et al. 2025).

Recent advancements have integrated LLMs into IE through two primary strategies: direct extraction and pre-trained language model (PLM)-assisted extraction. Direct extraction has evolved from supervised fine-tuning to generative paradigms, exemplified by UIE's unified text-to-structure framework (Zhao, Wang, and Kang 2022) and InstructUIE's multi-task instruction tuning (Wang et al. 2023; Liao et al. 2025; Li et al. 2025). PLM-assisted methods adopt hybrid architectures where LLMs either serve as primary extractors with PLMs for retrieval/calibration (Li et al. 2023; Zhang et al. 2024), or where PLMs perform extraction while LLMs generate synthetic training data (Zaratiana et al. 2024; Xu et al. 2023). This synergy has also spurred novel hybrid evaluation protocols (Fan et al. 2024).

Unlike existing approaches that treat fine-tuning and in-context learning as separate paradigms, our SCIR framework uniquely unifies these mechanisms within an LLM-assisted architecture, demonstrating superior performance through rigorous experimental validation.

Methodology

We first introduce the proposed SCIR framework, and then detail the construction of the MBSC dataset.

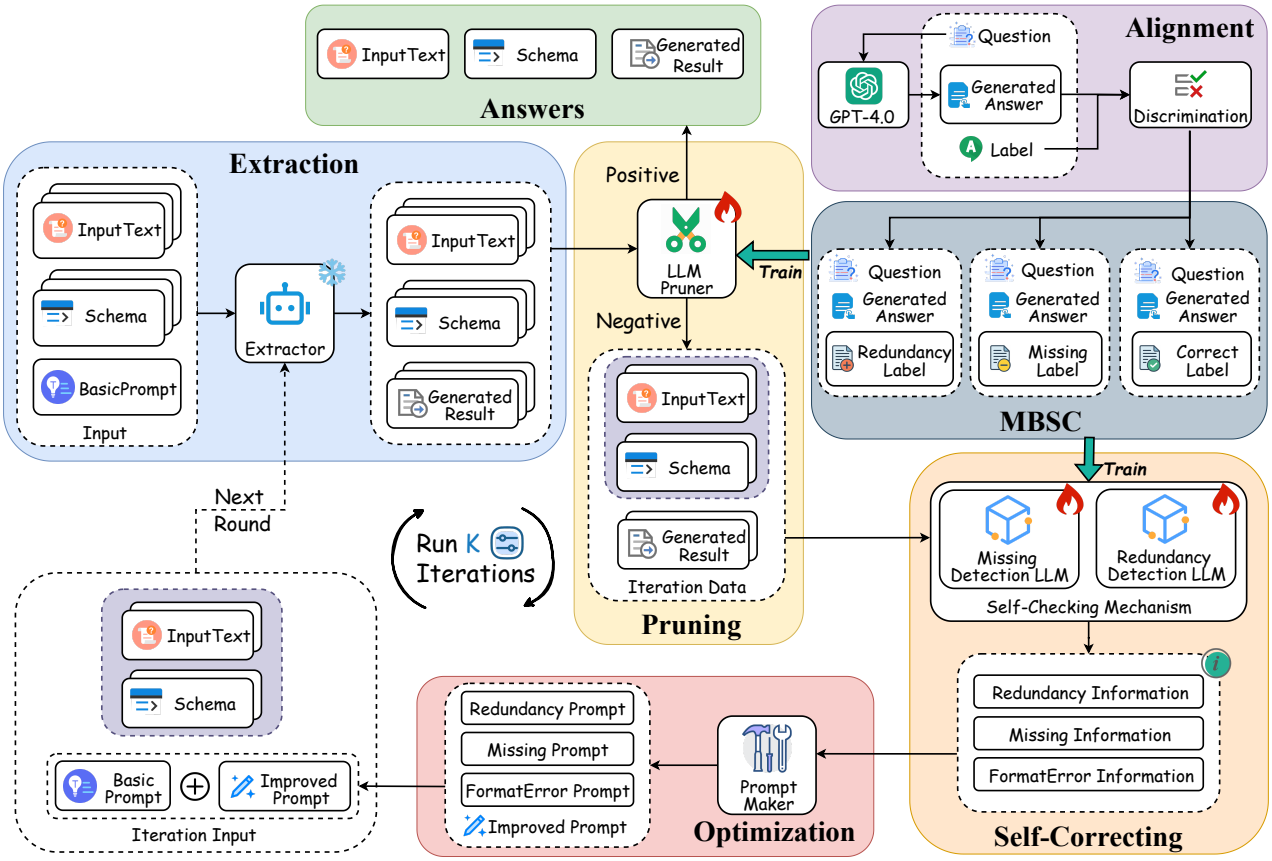

Figure 2: The architecture of SCIR, the number of iterations K is a hyperparameter.

SCIR Framework

As depicted in Figure 2, the SCIR framework enhances extraction quality through iterative refinement. The architecture consists of four core components: (1) an information extraction module, (2) an iterative pruning module, (3) a dual-path self-correcting module, and (4) a feedback-driven optimization module. These components operate in a synergistic pipeline to progressively improve extraction results through coordinated interactions. Below, we provide detailed descriptions of each module's structural design and functional mechanisms. The corresponding algorithmic implementation is detailed in Algorithm 1.

Information Extraction Module The Information Extraction Module is shown as the Extraction block in Figure 2. This module introduces a flexible paradigm distinct from conventional methods requiring domain-specific fine-tuning, supporting three extractor configurations: (1) untrained LLMs (Yang et al. 2025; DeepSeek-AI 2025; Meta AI 2024); (2) domain-adapted fine-tuned variants (Luo et al. 2024); and (3) existing IE frameworks (Liao et al. 2025). Initial extraction performs preliminary information extraction on raw text using basic instruction templates, while subsequent iterations dynamically refine results through optimized prompts filtered by the detection module. This design achieves dual advantages: plug-and-play compatibility with new models via standardized interfaces, enabling rapid integration without architectural modifications, and performance enhancement of existing models through in-context learning capabilities that iteratively guide the extraction process with refined prompts.

Result Pruning Module The Result Pruning Module addresses the efficiency challenge in iterative extraction by strategically identifying correct results for early termination. Recognizing that raw extraction outputs may already contain valid answers, we designed a discriminative pruning mechanism to bypass unnecessary iterations for confirmed-correct data. As depicted in the Pruning block of Figure 2, this component employs a Qwen3-4B (Yang et al. 2025)-based classifier trained on the MBSC dataset to partition extraction results into two categories: Positive samples meeting confidence thresholds are immediately output as final results, while Negative samples with potential errors are routed to the Dual-Path Self-Correcting module for refinement. This binary classification method effectively reduces computational load via early termination while ensuring accuracy by rectifying ambiguous results.

Dual-Path Self-Correcting Module The Dual-Path Self-Correcting Module enhances iterative extraction by simultaneously resolving redundancy and omission issues through a

Algorithm 1: Overall workflow of SCIR

Input: $Max\ Iterations: K, Basic\ Prompt: B_{prompt}, Data$
Output: $Answer_{set}$

$Answer_{set} = \{\}$
$Round_{prompt} = B_{prompt}$
$round = 0$
while $Data \neq \emptyset\ and\ round \leq K$ **do**
  $Gen_{result} \leftarrow LLM(Round_{prompt} \oplus Data)$
  $Check_{pos}, Check_{neg} \leftarrow Prun_{model}(Gen_{result})$
  if $Check_{pos} \neq \emptyset$ **then**
    $Answer_{set} \leftarrow Answer_{set} \cup Check_{pos}$
    $Data \leftarrow Data - Check_{pos}$
  end if
  if $Check_{neg} \neq \emptyset$ **then**
    $Red_{set} \leftarrow Red_{model}(Check_{neg})$
    $Mis_{set} \leftarrow Mis_{model}(Check_{neg})$
    $Red_{prompt} \leftarrow RP_{maker}(Red_{set})$
    $Mis_{prompt} \leftarrow MP_{maker}(Mis_{set})$
  end if
  $Iteration_{prompt} \leftarrow Red_{prompt} \cup Mis_{prompt}$
  $Round_{prompt} \leftarrow B_{prompt} \oplus Iteration_{prompt}$
  $round \leftarrow round + 1$
end while
if $round = K$ **then**
  $Answer_{set} \leftarrow Answer_{set} \cup Check_{neg}$
end if
return $Answer_{set}$

joint detection architecture. As illustrated in Figure 2's Self-Correcting block, this system employs two parallel paths: (1) **Redundancy Detection Path** systematically analyzes extraction outputs to identify and aggregate Redundancy Information into a structured Redundancy set, while (2) **Missing Detection Path** verifies logical and contextual coherence, generating a missing set. Both paths are also Qwen3-4B models fine-tuned by MBSC dataset. Additionally, any format violations detected during analysis are compiled into a FormatError set. Redundancy, Missing and FormatError sets provide multi-dimensional correction signals that enable precise error localization while maintaining full interpretability of the optimization process. The dual-path design achieves synergistic effects: the redundancy path ensures output conciseness, and the missing path guarantees completeness, collectively enhancing extraction quality through interpretable iterative refinement.

Feedback-Driven Optimization The Feedback-Driven Optimization module implements a closed-loop refinement system by injecting detection results into adaptive prompts. As shown in Figure 2, three diagnostic feedback streams generate specialized prompts: Redundancy Prompt, Missing Prompt, and FormatError Prompt. These prompts are dynamically fused with Basic Prompt into composite prompts, while preserving contextual semantics that guide LLM iterations, with each cycle incorporating updated feedback for progressive quality improvement.

MBSC Dataset

Traditional IE models rely on manually cleaned data for fine-tuning, which makes it difficult to effectively align with the model's prediction preferences. To align with the preferences of the Pruning module and the Self-Correcting module with the model's preferences, we created the MBSC dataset for training the model's detection capability and preference alignment. As shown in the Alignment block of Figure 2, built upon the IEPile dataset (Gui et al. 2024), MBSC employs GPT-4 to generate predictions for each IE question and systematically identifies three discrepancy patterns by comparing to original labels: missing information, redundant content, and correct matches. We employ GPT-4 as the generative model, as it represents the state-of-the-art in LLMs. Its characteristic errors often reflect common limitations across comparable models. For missing/redundant cases, we augment the label with absent content and a missing/redundant marker. For correct matches, we apply the *<Correct>* tag. This model behavior-driven approach constructs a training set that precisely evaluates content completeness while capturing common generation flaws even in advanced models like GPT-4. Unlike traditional synthetic methods relying on random modifications, MBSC's targeted label reconstruction mechanism achieves deeper alignment between detection capabilities and generative preferences, substantially improving the Self-Checking Mechanism's robustness through exposure to authentic error patterns from state-of-the-art models.

Dataset	Lang.	Task	Domain
COAE2016	ZH	RE	Web Text
SKE2020	ZH	RE	Commercial
Wiki-ZSL	EN	RE	Encyclopedia
FewRel (Han et al. 2018)	EN	RE	Experiment
Boson	ZH	NER	Financial
Weibo	ZH	NER	Social Media
CrossNER (Liu et al. 2021)	EN	NER	Experiment
CCF Law	ZH	EE	Legal
FewFC (Sheng et al. 2021)	ZH	EE	Judicial
RAMS	EN	EE	News
WikiEvents	EN	EE	Encyclopedia

Table 1: Dataset details.

Experiments

Experimental Setup

Datasets & Metrics To comprehensively validate our approach, we selected 11 benchmark datasets spanning three core IE tasks: (1) **Event Extraction**: CCF Law, FewFC, RAMS, and WikiEvents; (2) **Named Entity Recognition**: Boson, Weibo, and CrossNER; and (3) **Relation Extraction**: COAE2016, SKE2020, and FewRel. Dataset details are summarized in Table 1. Crucially, all experiments adhered to a strict zero-shot evaluation protocol where test sets were entirely excluded from training data. For quantitative assessment, we adopted the span-based Micro-F1 metric as the primary evaluation criterion. Based on the iterative round experiment results, we set the number of iterations to 2, achieving optimal efficiency without compromising effectiveness, and all results are based on this iteration round.

	EE			NER			RE		
Model	CCF Law	FewFC	Avg	Weibo	Boson	Avg	COAE2016	SKE2020	Avg
LLama3.1	31.57	32.52	32.05	17.02	29.74	23.38	28.66	34.74	31.70
Qwen3	34.99	41.29	38.14	19.01	35.42	27.21	25.49	36.73	31.11
DeepSeek-R1	38.79	40.81	39.80	24.66	40.67	32.67	32.69	45.32	39.00
YAYI-UIE	12.87	81.28	47.08	36.46	49.25	42.86	19.97	70.8	45.39
IEPile-LLama2	59.90	70.10	65.00	34.97	54.45	44.71	46.70	72.18	59.44
ChunkUIE	61.41	79.75	70.58	35.11	59.00	47.06	48.20	70.91	59.56
OneKE	62.19	80.11	71.15	35.06	**72.61**	53.84	49.83	72.61	61.22
SCIR-LLama3.1	60.99	63.75	62.37	31.21	58.03	44.62	**54.26**	67.68	60.97
SCIR-Qwen3	65.86	77.42	71.64	34.11	66.08	50.10	46.96	66.51	56.74
SCIR-DeepSeek-R1	62.45	68.20	65.32	39.68	68.45	54.07	52.45	**74.48**	**63.47**
SCIR-OneKE	**67.01**	**85.10**	**76.05**	**41.35**	68.58	**54.97**	51.05	73.81	62.43

Table 2: Performance comparison in Chinese IE Tasks. Best results are in bold and the second best are underlined.

	EE			NER	RE		
Model	RAMS	WikiEvents	Avg	CrossNER	Wiki-ZSL	FewRel	Avg
LLama3.1	10.26	7.13	8.69	26.65	13.65	19.14	16.39
Qwen3	12.67	8.27	10.47	30.90	16.01	17.75	16.88
DeepSeek-R1	13.02	8.68	10.85	37.84	24.98	21.79	23.38
YAYI-UIE	18.87	10.97	14.92	50.39	41.07	36.09	38.58
IEPile-LLama2	23.62	13.93	18.78	56.50	36.18	37.14	36.66
ChunkUIE	19.71	8.67	14.19	58.13	32.23	35.76	33.99
OneKE	22.58	12.43	17.51	60.91	42.18	39.19	40.69
RUIE	26.06	40.64	33.35	65.41	53.16	49.93	51.55
SCIR-LLama3.1	20.53	13.68	17.11	54.71	28.01	38.60	33.31
SCIR-Qwen3	24.70	15.61	20.15	61.51	31.64	34.67	33.17
SCIR-DeepSeek-R1	21.21	13.54	17.37	62.67	42.34	36.55	39.43
SCIR-OneKE	**27.04**	16.97	22.00	63.70	43.41	45.16	44.29
SCIR-RUIE	26.94	**45.74**	**36.34**	**65.54**	**53.71**	**55.02**	**54.37**

Table 3: Performance comparison in English IE Tasks. Best results are in bold and the second best are underlined.

Baselines We compare SCIR with several representative baselines, including untuned LLMs (LLama3.1, Qwen3, DeepSeek-R1), domain-specific models (YAYI-UIE, IEPile-LLama2, ChunkUIE and OneKE) and current mainstream IE frameworks (RUIE):

- **LLama3.1-8B** (Meta AI 2024)[1]: An open-source multilingual LLM released by Meta, which is suitable for multilingual conversations and text generation tasks.
- **Qwen3-8B** (Yang et al. 2025)[2]: Alibaba's 8-billion parameter dense model featuring RL-optimized performance in STEM and coding domains.
- **DeepSeek-R1-Distill-Qwen3-8B** (DeepSeek-AI 2025) [3]: This model is developed by DeepSeek through Chain-of-Thought distillation from its flagship model DeepSeek-R1-0528.

  YAYI-UIE (Xiao et al. 2023)[4]: A unified IE system trained on over 1 million human-annotated samples, supporting structured extraction across 12 distinct domains.
- **IEPile-LLama2** (Gui et al. 2024)[5]: A LLaMA2-13B model fine-tuned using LoRA on the IEPile dataset, demonstrating robust bilingual IE capabilities.
- **ChunkUIE** (Li et al. 2025)[6]: Implements chunked instruction processing and hard negative sampling to address semantic ambiguity in bilingual IE tasks.
- **OneKE** (Luo et al. 2024)[7]: A large-scale model IE framework featuring bilingual support and generalization capabilities for multiple domains and tasks.
- **RUIE** (Liao et al. 2025)[8]: An English-only extraction framework utilizing BM25 sparse retrieval for efficient candidate screening.

Moreover, to assess SCIR's plug-and-play capability, we implemented two sets of its variants for comparison: 1) SCIR based on untuned LLMs (SCIR-LLama3.1, SCIR-Qwen3 and SCIR-DeepSeek-R1), and 2) SCIR integrated with representative domain-specific models (SCIR-OneKE) and mainstream IE frameworks (SCIR-RUIE).

Main Results

Chinese Dataset Performance As presented in Table 2, the SCIR framework demonstrates outstanding performance

[1]https://github.com/meta-llama/llama3.
[2]https://github.com/QwenLM/Qwen3.
[3]https://github.com/deepseek-ai/DeepSeek-R1.
[4]https://github.com/wenge-research/YAYI-UIE
[5]https://github.com/zjunlp/IEPile.
[6]https://github.com/ChunkUIE/chunkuie.
[7]https://github.com/zjunlp/OneKE.
[8]https://github.com/OStars/RUIE.

Task	Dataset	w/o Both	w/o Red	w/o Mis	FULL
EE	CCF Law	34.99	63.62	62.73	**65.86**
	FewFC	41.29	70.12	74.44	**77.42**
	RAMS	12.67	21.79	22.71	**24.70**
	WikiEvents	8.27	11.14	13.32	**15.61**
NER	boson	35.42	64.45	61.27	**66.08**
	WEIBONER	19.01	31.95	31.79	**34.11**
	CrossNER	30.90	57.05	58.39	**61.51**
RE	COAE2016	25.49	38.55	42.40	**46.96**
	SKE2020	36.73	59.52	62.23	**66.51**
	Wiki-ZSL	16.01	28.45	27.48	**31.64**
	FewRel	17.75	33.11	28.77	**34.67**

Table 4: Module ablation study results, where 'w/o Red' denotes using only the missing detection module and 'w/o Mis' denotes using only the redundant detection module.

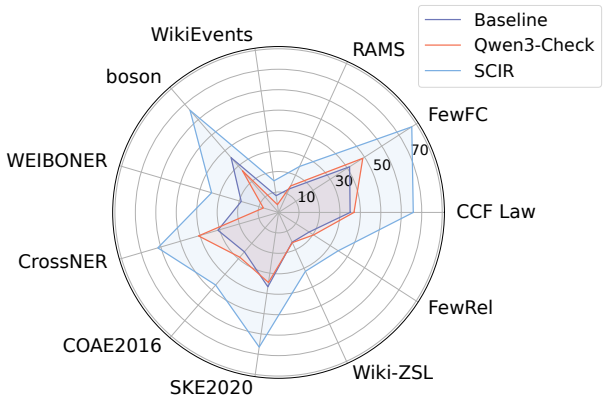

Figure 3: The figure shows the F1 scores of the three model ablation experiments on each dataset. The term "Baseline" refers to only use the LLMs to extraction, while "Qwen3-Check" denotes using an untuned Qwen3-4B model in Dual-Path Self-Correcting Module.

on Chinese datasets. In EE tasks, SCIR achieves exceptional results when integrated with vertical domain-specific models, while showing slightly reduced performance when combined with base LLMs—likely due to prompt comprehension challenges caused by event structure complexity. Nevertheless, both integration approaches significantly outperform baseline systems. For NER tasks, SCIR's performance improvement appears relatively modest, potentially constrained by the inherent simplicity of the task. In RE tasks, SCIR excels particularly well, substantially enhancing the typically mediocre performance of LLMs in this task. We attribute this improvement to the Dual-Path Self-Correcting mechanism's precise control over redundant and missing content, effectively unleashing the potential of LLMs. Notably, SCIR consistently delivers significant performance gains across different extractors, demonstrating its excellent plug-and-play capability.

English Dataset Performance As shown in Table 3, SCIR maintains excellent performance on English datasets. Benefiting from the advantages of the robust RUIE framework, SCIR achieves or approaches state-of-the-art performance across all three IE tasks. However, compared to its performance on Chinese datasets, SCIR demonstrates slightly diminished performance gains on English datasets. We attribute this primarily to the fact that the Dual-Path Self-Correcting mechanism was trained using the Qwen3-4B model, which inherently possesses stronger capabilities for Chinese tasks due to its corpus composition. Notably, SCIR successfully achieves effective integration with base LLMs, domain-specific models, and IE frameworks, fully demonstrating its superior model generalization capability.

In sum, the SCIR framework exhibits substantial performance gains across Chinese and English datasets, delivering a 5.27% improvement over current state-of-the-art bilingual extraction models (OneKE). For Chinese tasks, SCIR's Dual-Path Self-Correcting mechanism effectively augments base LLMs (Qwen3, LLaMA-3.1, DeepSeek) on complex tasks while maintaining seamless integration with domain-specific models like OneKE. For English tasks, SCIR successfully combines with existing frameworks (e.g., RUIE) to achieve cutting-edge performance. These comprehensive cross-lingual evaluations demonstrate SCIR's remarkable flexibility and compatibility, enabling true plug-and-play functionality with diverse model architectures.

Ablation Study

Module Ablation Study In this set of experiments, we focus on the 'Dual-Path Self-Correcting Module' since other components are essential or performance-neutral. Our experiments compare three configurations: redundant-only, missing-only, and full dual-path detection. Table 4 shows the complete implementation achieves statistically significant gains across all tasks, with distinct patterns emerging across different domains. In EE, redundant detection proves more impactful, while missing detection shows stronger effects in NER. For RE, we observe language-dependent variations - redundant detection excels on Chinese datasets whereas missing detection performs better on English corpora. These findings conclusively validate the dual-path design, demonstrating how the two modules operate in complementary fashion to enhance overall model performance.

Model Ablation Study The model ablation study compares the untrained Qwen3-4B model as the experimental group with the same model trained on the MBSC dataset to demonstrate that SCIR's performance improvement stems from the Dual-Path Self-Correcting mechanism rather than the base model itself. The results in Figure 3 show that the performance gains from the untrained Qwen3-4B model are negligible, while the same model trained on the MBSC dataset significantly enhances IE performance. Notably, when using the untrained Qwen3-4B model, performance degradation is observed across the WikiEvent, boson, and SKE2020 datasets, indicating that model errors are amplified iteratively due to the lack of timely correction. The ablation study confirms that SCIR's performance improvement originates from the Dual-Path Self-Correcting mechanism rather than the base model itself, while also highlighting the importance of the MBSC dataset.

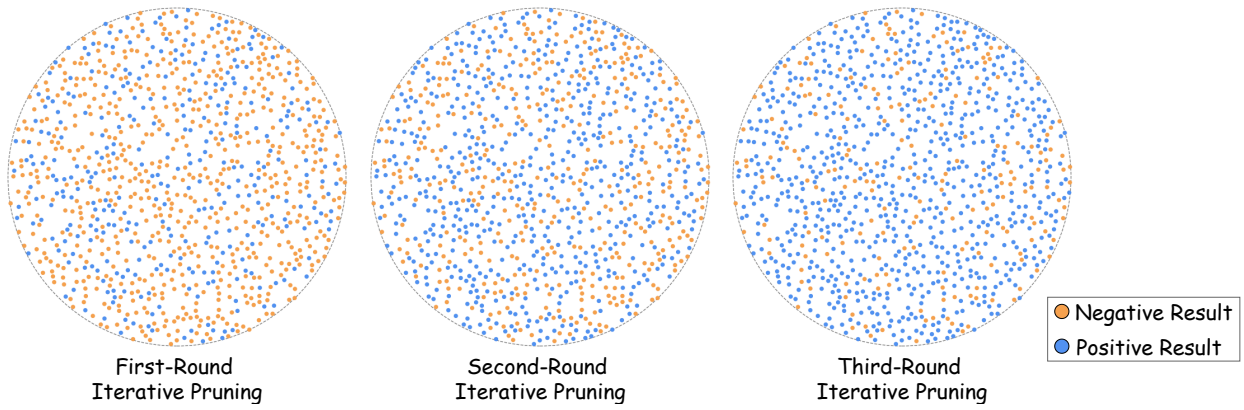

Figure 4: Iterative Experimental Results.

Task	Time	Performance
Event Extraction (EE)	+13.72%	+48.21%
Named Entity Recognition (NER)	+11.96%	+42.66%
Relation Extraction (RE)	+14.32%	+43.68%
Average	+13.33%	+44.86%

Table 5: The table presents the percentage of generation time occupied by SCIR framework's iterative detection and the performance improvement achieved using the SCIR.

Validation of Pruning Efficacy

Through statistical analysis of the iterative pruning performance of SCIR-Deepseek-R1 on the SKE2020 dataset, as shown in Figure 4, we validate the effectiveness of the pruning module. Each point in the figure represents a data item to be extracted, with orange dots indicating instances that were either incorrectly retained or correctly pruned, and blue dots denoting correctly pruned instances. Notably, as iterations progress, the number of correctly pruned points increases significantly. This observation demonstrates both the pruner's practical efficacy and its ability to generate more precise pruning decisions for subsequent iterations based on prior results. These findings robustly confirm the necessity of the pruner and verify the absence of error propagation or amplification throughout the iterative process.

Experiment on Time Costs

We have meticulously recorded the time costs incurred during both training and inference. (1) In terms of training, SCIR attains convergence within 3 hours when utilizing 4 RTX4090 GPUs. By contrast, under the same hardware, traditional methods for training vertical-domain models necessitate 22 hours, translating to an approximate 87% reduction in time cost. (2) Regarding inference, as presented in Table 5, we compare the average time consumption and the corresponding average performance enhancements of SCIR when integrated with DeepSeek-R1, LLama3.1, Qwen3, OneKE, and RUIE across three tasks. The results reveal that, by harnessing efficient pruning techniques and the swift inference capabilities of lightweight detectors, SCIR achieves remarkable performance improvements while introducing only a marginal increase in time cost overhead.

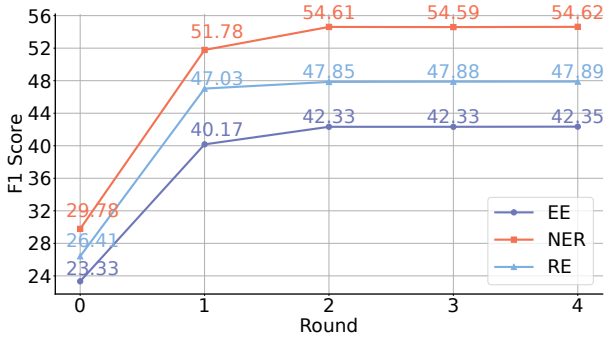

Figure 5: Average F1 scores of the SCIR framework across multiple iterations for three IE tasks.

Experiment on Iterative Round

We statistically analyzed the average F1 scores of SCIR across multiple iterations for three IE tasks, as depicted in Figure 5. The results indicate that SCIR achieves significant performance gains in the first two iterations, with diminishing improvements thereafter. Consequently, we set the iteration limit to 2 for all experiments. Notably, even a single iteration yields substantial performance enhancements, strongly validating SCIR's optimization effectiveness.

Conclusion

This study proposes the Self-Correcting Iterative Refinement (SCIR) framework, whose effectiveness in enhancing IE performance has been comprehensively validated across 11 bilingual datasets covering diverse tasks. The framework demonstrates three core advantages: superior extraction accuracy, effective model preference alignment, and low-cost model portability. These characteristics position SCIR as an innovative solution for developing lightweight and reusable IE systems, while providing a new reference paradigm for future research in the field of information extraction.

References

Chiticariu, L.; Li, Y.; and Reiss, F. R. 2013. Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of EMNLP*, 827–832. Seattle, Washington, USA: Association for Computational Linguistics.

Cortes, C.; and Vapnik, V. 1995. Support-Vector Networks. *Mach. Learn.*, 20(3): 273–297.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186.

Fan, Y.; Liu, Y.; Yao, Z.; Yu, J.; Hou, L.; and Li, J. 2024. Evaluating Generative Language Models in Information Extraction as Subjective Question Correction. In Calzolari, N.; Kan, M.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 6409–6417.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR*, abs/2312.00752.

Gui, H.; Yuan, L.; Ye, H.; Zhang, N.; Sun, M.; Liang, L.; and Chen, H. 2024. IEPile: Unearthing Large Scale Schema-Conditioned Information Extraction Corpus. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, 127–146.

Han, F.; Yu, X.; Tang, J.; and Ungar, L. H. 2025. ZeroTuning: Unlocking the Initial Token's Power to Enhance Large Language Models Without Training. *CoRR*, abs/2505.11739.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 4803–4809.

Jiang, C.; Chan, C.; Xue, W.; Liu, Q.; and Guo, Y. 2025. Importance Weighting Can Help Large Language Models Self-Improve. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 24257–24265.

Li, M.; Chen, M.; Zhou, H.; and Zhang, R. 2023. PeTailor: Improving Large Language Model by Tailored Chunk Scorer in Biomedical Triple Extraction. *CoRR*, abs/2310.18463.

Li, W.; Liu, Y.; Yang, Y.; Zhang, T.; and Men, W. 2025. ChunkUIE: Chunked instruction-based unified information extraction. *PLOS ONE*, 20(6): e0326764.

Liao, X.; Duan, J.; Huang, Y.; and Wang, J. 2025. RUIE: Retrieval-based Unified Information Extraction using Large Language Model. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, 9640–9655.

Liu, Z.; Xu, Y.; Yu, T.; Dai, W.; Ji, Z.; Cahyawijaya, S.; Madotto, A.; and Fung, P. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 13452–13460.

Lou, J.; Lu, Y.; Dai, D.; Jia, W.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2023. Universal Information Extraction as Unified Semantic Matching. In Williams, B.; Chen, Y.; and Neville, J., eds., *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*, 13318–13326.

Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 2022. Unified Structure Generation for Universal Information Extraction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 5755–5772.

Luo, Y.; Ru, X.; Liu, K.; Yuan, L.; Sun, M.; Zhang, N.; Liang, L.; Zhang, Z.; Zhou, J.; Wei, L.; Zheng, D.; Wang, H.; and Chen, H. 2024. OneKE: A Dockerized Schema-Guided LLM Agent-based Knowledge Extraction System. *CoRR*, abs/2412.20005.

Ma, Y.; Cao, Y.; Hong, Y.; and Sun, A. 2023. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 10572–10601.

Mao, Y.; Ge, Y.; Fan, Y.; Xu, W.; Mi, Y.; Hu, Z.; and Gao, Y. 2025. A survey on LoRA of large language models. *Frontiers Comput. Sci.*, 19(7): 197605.

Maturana, F.; Riveros, C.; and Vrgoč, D. 2017. Document Spanners for Extracting Incomplete Information: Expressiveness and Complexity. *arXiv preprint arXiv:1707.00827*.

Meta AI. 2024. Llama 3.1 Technical Report: Language, Vision, and Multimodal Capabilities. Technical report, Meta. Model parameters: 405B, context length 128K.

Qi, Y.; Peng, H.; Wang, X.; Xu, B.; Hou, L.; and Li, J. 2024. ADELIE: Aligning Large Language Models on Information Extraction. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 7371–7387.

Sheng, J.; Guo, S.; Yu, B.; Li, Q.; Hei, Y.; Wang, L.; Liu, T.; and Xu, H. 2021. CasEE: A Joint Learning Framework with Cascade Decoding for Overlapping Event Extraction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 164–174.

Tedeschi, S.; and Navigli, R. 2022. MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation). In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, 801–812.

Thenmozhi, D.; and Aravindan, C. 2018. RCE-OIE: Open Information Extraction Using a Rule-Based Clause Extraction Engine. In *Recent Findings in Intelligent Computing Techniques*, volume 709 of *Advances in Intelligent Systems and Computing*, 191–198. Springer.

Wan, F.; Huang, X.; Cai, D.; Quan, X.; Bi, W.; and Shi, S. 2024. Knowledge Fusion of Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; Wang, G.; and Guo, C. 2025. GPT-NER: Named Entity Recognition via Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 4257–4275.

Wang, X.; Zhou, W.; Zu, C.; Xia, H.; Chen, T.; Zhang, Y.; Zheng, R.; Ye, J.; Zhang, Q.; Gui, T.; Kang, J.; Yang, J.; Li, S.; and Du, C. 2023. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. *CoRR*, abs/2304.08085.

Wilks, Y. 1997. Information Extraction as a Core Language Technology. In Pazienza, M. T., ed., *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, SCIE-97, Frascati, Italy, 14-18, 1997*, volume 1299 of *Lecture Notes in Computer Science*, 1–9.

Xiao, X.; Wang, Y.; Xu, N.; Wang, Y.; Yang, H.; Wang, M.; Luo, Y.; Wang, L.; Mao, W.; and Zeng, D. 2023. YAYI-UIE: A Chat-Enhanced Instruction Tuning Framework for Universal Information Extraction. *CoRR*, abs/2312.15548.

Xiong, T.; Wei, W.; Xu, K.; and Chen, D. 2025. SA-DETR: Span Aware Detection Transformer for Moment Retrieval. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, 7634–7647.

Xu, B.; Wang, Q.; Lyu, Y.; Dai, D.; Zhang, Y.; and Mao, Z. 2023. S2ynRE: Two-stage Self-training with Synthetic data for Low-resource Relation Extraction. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 8186–8207.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yang, Y.; Tang, Y.; and Tam, K. Y. 2023. InvestLM: A Large Language Model for Investment using Financial Domain Instruction Tuning. *CoRR*, abs/2309.13064.

Zaratiana, U.; Tomeh, N.; Holat, P.; and Charnois, T. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 5364–5376.

Zhang, X. F.; Blum, C. W.; Choji, T.; Shah, S.; and Vempala, A. 2024. ULTRA: Unleash LLMs' Potential for Event Argument Extraction through Hierarchical Modeling and Pairwise Self-Refinement. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 8172–8185.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhao, F.; Wang, Y.; and Kang, Y. 2022. A Prompt-Based UIE Framework. In Zhang, N.; Wang, M.; Wu, T.; Hu, W.; and Deng, S., eds., *CCKS 2022 - Evaluation Track - 7th China Conference on Knowledge Graph and Semantic Computing Evaluations, CCKS 2022, Qinhuangdao, China, August 24-27, 2022, Revised Selected Papers*, volume 1711 of *Communications in Computer and Information Science*, 163–171.

Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; and Xu, B. 2017. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1227–1236.