IEPILE: Unearthing Large-Scale Schema-Based Information Extraction Corpus

Honghao Gui[♠♣♢*], **Lin Yuan**[♣♢*], **Hongbin Ye**[♠], **Ningyu Zhang**[♠♢†]
Mengshu Sun[♣♢], **Lei Liang**[♣♢], **Huajun Chen**[♠♢†]
[♠] Zhejiang University [♣] Ant Group
[♢] Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph
{guihonghao,zhangningyu,huajunsir}@zju.edu.cn

https://github.com/zjunlp/IEPile

Abstract

Large Language Models (LLMs) demonstrate remarkable potential across various domains; however, they exhibit a significant performance gap in Information Extraction (IE). Note that high-quality instruction data is the vital key for enhancing the specific capabilities of LLMs, while current IE datasets tend to be small in scale, fragmented, and lack standardized schema. To this end, we introduce IEPILE, a comprehensive bilingual (English and Chinese) IE instruction corpus, which contains approximately **0.32B** tokens. We construct IEPILE by collecting and cleaning 33 existing IE datasets, and introduce schema-based instruction generation to unearth a large-scale corpus. Experimentally, IEPILE enhance the performance of LLMs for IE, with notable improvements in zero-shot generalization. We open-source the resource and pre-trained models, hoping to provide valuable support to the NLP community.

1 Introduction

Large Language Models (LLMs) have achieved significant breakthroughs in multiple Natural Language Processing (NLP) tasks (Du et al., 2022; Touvron et al., 2023b; Jiang et al., 2023; Zhao et al., 2023; Pu et al., 2023; Yang et al., 2024; Wu et al., 2023; Wang et al., 2023c; Fei et al., 2024). However, recent studies (Li et al., 2023a; Ma et al., 2023; Xu et al., 2023; Wadhwa et al., 2023; Wan et al., 2023; Gao et al., 2023; Li et al., 2023b; Jiao et al., 2023; Huang et al., 2023; Wang et al., 2024) indicate a significant performance gap in the task of Information Extraction (IE) when utilizing LLMs. (Lee et al., 2022a; Gao et al., 2023) further illustrate that the major reason may lie in limited high-quality, large-scale data corpus. Concretely, most IE datasets are often limited in size, scattered in distribution, and lack standardization in schema[1].

Faced with these limitations, there is an urgent need to collect instruction data in a unified and automated manner to build a high-quality, large-scale IE corpus. To this end, we collect and clean various existing IE datasets to obtain a comprehensive bilingual IE instruction dataset named IEPILE[2]. During the corpus construction, we find existing methods for constructing IE instruction data suffer from two issues for generalizable IE: 1) **Schema Query Disparity**: There may be inconsistency in the number of schema queries within instruction between training and evaluation which can harm model generalization; 2) **Semantic Confusion**: The co-occurrence of semantically similar schemas within instructions may confuse the model. Thus, we introduce a schema-based instruction generation strategy. We first construct a hard negative schema dictionary to promote the more frequent occurrence of semantically similar schema in instructions. Then, we introduce batched instruction generation, dynamically limiting the number of schemas queried in each instruction to $split\_num$, which not only addresses the issue of performance degradation due to inconsistent numbers of schema queries during training and evaluation, but also enhances the robustness when dealing with semantically confusing schema. Finally, we obtain IEPILE which contains approximately 0.32B tokens.

By fine-tuning a selection of the latest prominent models (Yang et al., 2023; Touvron et al., 2023b; Bai et al., 2023) on the IEPILE dataset, we show that LLMs with IEPILE can yield better zero-shot performance than baselines. This achievement not only verifies the effectiveness of the IEPILE dataset but also provides a framework for creating IE datasets in other domains.

* Equal Contribution.
† Corresponding Author.

[1] We refer to the schema as pre-defined types of entities, relations, events (arguments and roles), etc.

[2] IEPILE adhere to the CC BY-NC-SA 4.0 license except for ACE2005 which adheres to the LDC User Agreement.

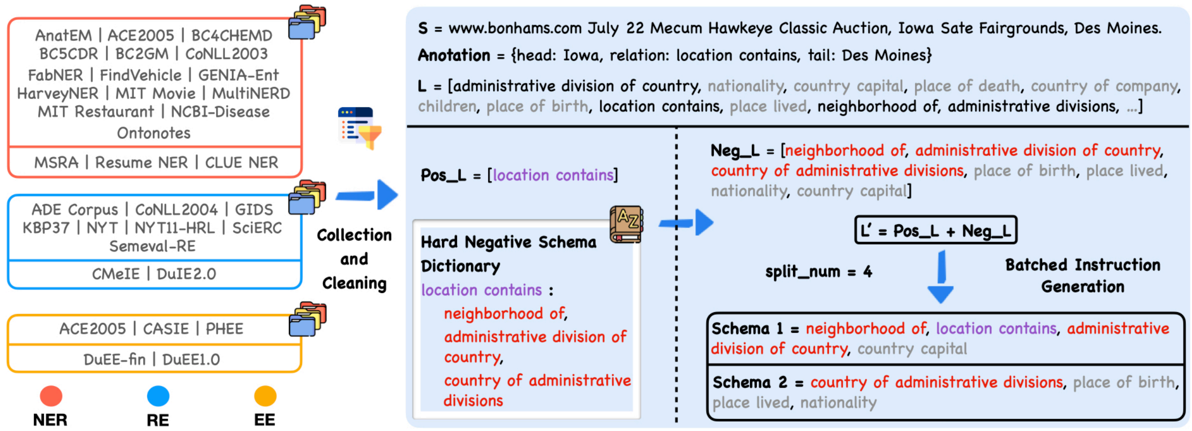

Figure 1: An overview of the construction of IEPILE, including Data Collection and Cleaning, as well as Schema-Based Instruction Generation (Hard Negative Schema Construction and Batched Instruction Generation).

2 IEPILE

In this section, we introduce the construction of IEPILE and provide details in Appendix B.

2.1 Data Collection and Cleaning

To broadly cover various domains and meet the practical demands, we collect datasets necessary for IE from multiple data sources. Our corpus mainly involves bilingual data (Chinese and English) and focuses on three principal categories of IE tasks: Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE). In total, we gather 26 English datasets and 7 Chinese datasets. We also employ standardization procedures to maintain data quality and format uniformity, involving format unification, instance deduplication, and the exclusion of low-quality data.

2.2 Schema-Based Instruction Generation

We concentrate on instruction-based information extraction (IE), a methodology that incorporates three crucial elements to compose an instruction: 1) **Task Description**, a template utilized to distinguish between different IE tasks; 2) **Input Text**, the source text to be extracted; and 3) **Schema sequence**, which defines the information that the model is supposed to extract, including entity types, relations, events, etc. Among these, the schema sequence is critical as it reflects the specific extraction requirements and is dynamically variable. Therefore, the construction of the schema sequence within an instruction holds critical significance.

Positive and Negative Schema Mechanism in Instructions. Firstly, we define schemas that actually exist within the input text as **positive schemas** and those that do not appear as **negative schemas**. As illustrated in Figure 1, the "location contains" present in the annotation is a positive schema, while all other schemas from the predefined label set L are negative schemas. Traditional IE frameworks, which are treated as sequence labeling tasks, take text as input and produce a label for each token as output, without involving the concept of positive or negative schemas within the model's input. However, in the era of generative IE, represented by models like UIE (Lu et al., 2022a), introduce the concept of integrating a schema sequence (refers to as Structural Schema Instructor, or SSI) in the model's input to guide its output, restricting the range of output to the SSI. The method necessitates including the entire predefined label set of a dataset as the SSI to guide the model's output during inference. As a result, if the SSI during the training contains only positive schemas, the model will tend to generate corresponding answers for every label within the SSI during inference. Therefore, to make the model explicitly reject generating outputs for negative schemas, it is necessary to incorporate negative schemas into the SSI.

In this paper, the schema sequence included in the instructions follows the concept of SSI. However, we observe that existing research (Wang et al., 2023b; Xiao et al., 2023) tends to adopt a rather crude schema processing strategy when constructing instructions, meaning that all schemas within a predefined label set are used to build the instructions. This approach potentially entails two significant issues: 1) **Inconsistency in the number of schema queries within instruction between training and evaluation**. For example, the model's performance will decrease if it is trained on about

Algorithm 1 Schema-Based Instruction Generation

Require: Text S, Predefined label set L, Positive schema set Pos_L, Number of schemas to split $split\_num$
Ensure: Set of $Instructions$
  Step 1: Initialize Hard Negative Schema Dictionary $\mathcal{K}$
  for all schema in L **do**
    $\mathcal{K}[schema] \leftarrow$ SEMANTIC-SIMILAR$(schema, L)$
  end for
  Step 2: Obtain Hard Negative Schemas
  $Hard\_L \leftarrow \emptyset$
  for all schema in Pos_L **do**
    $Hard\_L \leftarrow Hard\_L \cup \mathcal{K}[schema]$
  end for
  $Other\_L \leftarrow L - Pos\_L - Hard\_L$
  $Other\_L \leftarrow$ RANDOM-SELECT$(Other_L, split_num)$

  $Neg\_L \leftarrow Hard\_L \cup Other\_L$
  $L' \leftarrow Neg\_L \cup Pos\_L$
  Shuffle L' to obtain a randomized sequence
  Step 3: Batched Instruction Generation
  $Instructions \leftarrow []$
  $num\_batches \leftarrow \lceil \frac{|L'|}{split\_num} \rceil$
  for $i \leftarrow 1$ **to** $num\_batches$ **do**
    $Batch \leftarrow$ SEQUENTIAL-SELECT$(L', split_num, i)$

    $Instructions \leftarrow Instructions \cup$ GENERATE-INSTRUCTION$(Batch)$
  end for

20 schema queries but tested with either 10 or 30, even if the training and evaluation schemas are similar in content. 2) **Inadequate differentiation among schemas in the instructions**. For example, semantically similar schemas like "layoffs", "depart" and "dismissals", may present co-occurrence ambiguities that could confuse the LLMs. Such schemas should co-occur more frequently within the instruction. Therefore, we introduce: 1) Hard Negative Schema Construction; and 2) Batched Instruction Generation. Detailed information can be found in Figure 1 and Algorithm 1.

Hard Negative Schema Construction. As illustrated in Figure 1, assume that dataset $\mathcal{D}$ possesses a predefined label set L. For a given text S, the schemas present in its annotation constitute the positive schema set Pos_L, while others form the negative schema set Neg_L. In our analysis, we discover that the primary cause of model mistakes stems from the semantic ambiguity of the schema. In traditional approaches, the Neg_L is simply defined as $L - Pos\_L$. However, they overlook a critical aspect: it is important to pay special attention to negative schemas that are semantically similar to positive schemas. Inspired by the theory of contrastive learning, we propose the concept of a hard negative schema dictionary $\mathcal{K}$, where each key represents a unique schema and

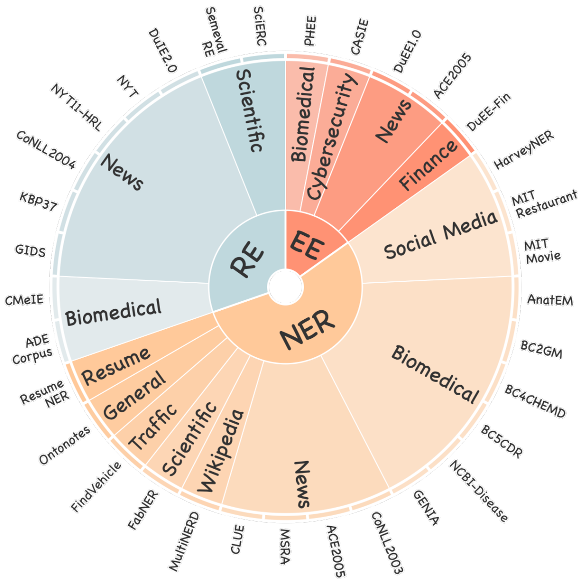

Figure 2: Distribution of different tasks, domains, and source datasets within the IEPILE.

the associated value is a collection of schemas that are semantically similar to the key schema. The hard negative schemas are constructed by querying GPT-4 and manually reviewing them. Based on this, we define the hard negative schema set as $Hard\_L = \mathcal{K}[Pos\_L]$, and the other negative schema set as $Other\_L = L - Pos\_L - Hard\_L$. The final Neg_L is constituted by $Hard\_L$ and a small subset of $Other\_L$. Through this strategy, we not only present semantically similar schemas more frequently within the instruction but also reduce the number of training instances without sacrificing model performance.

Batched Instruction Generation. Subsequently, we obtain the final schema set $L' = Pos\_L + Neg\_L$. We employ a batched instruction generation method, dynamically limiting the number of schemas inquired in each instruction to the number of $split\_num$, which ranges between 4 and 6. Therefore, L' will be divided into $|L'|/split\_num$ batches for querying, with each batch querying $split\_num$ schemas. Consequently, even if the number of schemas inquired during the evaluation phase differs from that of training, the batched mechanism allows us to distribute the inquiries across $split\_num$ schemas, thereby mitigating the decline in generalization performance.

2.3 Data Statistics

Based on the aforementioned methods, we obtain the IEPILE dataset, which includes roughly 2 million instruction entries and approximately 0.32B to-

Method	NER	RE			EE			
	CrossNER	FewRel	Wiki-ZSL	Avg	WikiEvents	RAMS	CrudeOil News	Avg
LLaMA2	34.82	6.53	9.43	7.98	0.00	0.00	0.00	0.00
Baichuan2	38.93	5.94	4.15	5.05	0.00	0.00	0.00	0.00
Qwen1.5	50.13	7.82	6.94	7.38	0.00	0.00	0.00	0.00
Mistral	42.83	6.84	5.10	5.97	0.00	0.00	0.00	0.00
ChatGPT	58.37	9.96	13.14	11.55	2.95	8.35	1.41	4.24
GPT-4	58.49	22.43	23.76	23.10	5.24	10.14	26.13	13.84
UIE	38.37	-	-	-	5.12	9.25	6.45	6.94
InstructUIE	49.36	39.55	35.20	37.38	11.64	**24.27**	23.26	19.72
YAYI-UIE	50.39	36.09	41.07	38.58	10.97	18.87	12.45	14.10
Baichuan2-IEPILE	55.55	**41.28**	37.61	39.45	9.12	20.19	36.61	21.97
LLaMA2-IEPILE	56.50	37.14	36.18	36.66	**13.93**	23.62	33.87	23.81
Qwen1.5-IEPILE	57.90	40.92	38.49	39.71	11.38	21.26	30.69	21.11
LLaMA3-IEPILE	56.11	35.58	37.18	36.38	9.71	20.27	**39.88**	23.29
OneKE	**60.91**	39.19	**42.18**	**40.68**	12.43	22.58	38.49	**24.50**

Table 1: Zero-shot performance on English datasets. UIE necessitates predefined entity types; given that such information is not provided by the FewRel and Wiki-ZSL datasets, we are unable to evaluate UIE's performance on these datasets. For the task of event extraction, we only present the results of event detection in the main text.

kens (utilizing the Baichuan2 tokenizer). Figure 2 displays the distribution of domains and source datasets within the IEPILE, including 33 datasets spanning multiple domains such as general, news, finance, and biomedical. Additionally, Table 12 provides examples of instructions and outputs for 3 different tasks within the IEPILE.

3 Experiments

Based on IEPILE, we fine-tune several latest prominent models, then compare their zero-shot generalization capabilities against a range of baseline models. Results of the full supervision evaluation and training details are described in Appendix C.

3.1 Experimental Settings

Evaluation Metrics: We employ span-based Micro-F1 as the metric for measuring model performance. **Baselines**: We select a range of strong models for comparative analysis, which include UIE (Lu et al., 2022a), LLaMA2-13B-Chat (Touvron et al., 2023b), Baichuan2-13B-Chat (Yang et al., 2023), Qwen1.5-14B-Chat (Bai et al., 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), ChatGPT (Ouyang et al., 2022), GPT-4 (OpenAI, 2023), LLaMA3-8B-Instruct, InstructUIE (Wang et al., 2023b), YAYI-UIE (Xiao et al., 2023). **Zero-shot Benchmark**: We collect 13 datasets that are **not present in the training set**. **OneKE**: Additionally, we perform full-parameter fine-tuning of the alpaca2-chinese-13B model utilizing IEPILE and other proprietary information extraction datasets. This paper also reports its results; for more detailed information, please refer to Appendix C.2.

3.2 Main Results

In Tables 1 and 2, we report the zero-shot performance across three tasks and two languages. Overall, after training with the IEPILE, the models achieve better results in the majority of tasks. We believe the success is due to the hard negative schema construction and batched instruction generation strategy, which can mitigate the train-eval mismatch and semantic ambiguity for the diverse schema. We also observe that IEPILE-models are slightly behind GPT-4 in English NER. We hypothesize that the marginal gap may be attributed to GPT-4's exposure to a vast corpus of similar data during its training. Moreover, it is essential to note that InstructUIE focuses on English data while IEPILE incorporates both English and Chinese data. This disparity in data may influence the capability of the model in English, potentially reducing the performance. Additionally, OneKE achieves the best results in nearly all zero-shot evaluation tasks. We attribute this success to the enhancements brought by full parameter fine-tuning.

3.3 Analysis

Inconsistency in the Number of Schema Queries Hurt Generalization. We investigate the impact on model performance when different numbers of schema queries are used during the training and

Method	NER			RE				EE		
	Boson	Weibo	Avg	SKE2020	COAE2016	IPRE	Avg	FewFC	CCF Law	Avg
LLaMA2	8.19	2.43	5.31	0.50	3.11	0.31	1.31	0.23	0.08	0.16
Baichuan2	27.39	7.62	17.51	7.23	11.65	1.45	6.78	11.82	2.73	7.28
Qwen1.5	26.49	25.34	25.92	7.69	11.97	2.16	7.27	11.47	3.25	7.36
Mistral	29.13	10.02	19.58	6.84	5.24	0.82	4.30	4.69	0.23	2.46
ChatGPT	38.53	29.30	33.92	24.47	19.31	6.73	16.84	16.15	0.00	8.08
GPT-4	48.15	29.80	38.98	56.77	41.15	18.15	38.69	74.25	42.12	58.19
YAYI-UIE	49.25	36.46	42.86	70.80	19.97	22.97	37.91	81.28	12.87	47.08
Baichuan2-IEPILE	55.77	**38.03**	46.90	72.50	47.43	29.76	49.90	**83.59**	**63.53**	**73.56**
LLaMA2-IEPILE	54.45	34.97	44.71	72.18	46.70	28.55	49.14	70.10	59.90	65.00
Qwen1.5-IEPILE	63.08	37.50	50.29	72.29	**50.70**	30.55	51.18	78.77	61.43	70.10
LLaMA3-IEPILE	61.88	37.43	49.66	73.67	48.12	**31.29**	51.03	81.52	59.92	70.72
OneKE	**72.61**	35.06	**53.84**	**74.15**	49.83	29.95	**51.31**	80.11	62.19	71.15

Table 2: Zero-shot performance on Chinese datasets. Since UIE and InstructUIE do not train with Chinese data, we do not report performance of these two models on Chinese datasets.

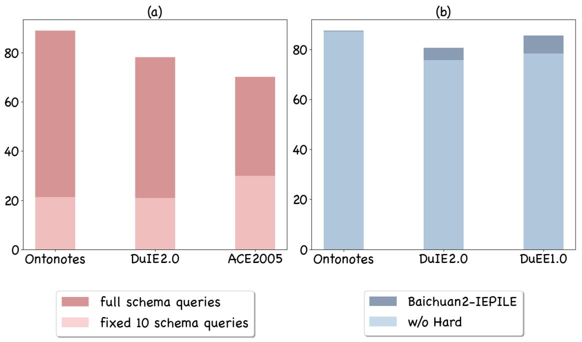

Figure 3: (a) When there is an inconsistency in the number of schema inquiries during the training and evaluation, the performance of the model significantly decreases. (b) The impact of removing the hard negative schema dictionary on the performance of the model.

evaluation. We train the Baichuan2 using full-schema instructions on 3 datasets: Ontonotes (18 schemas), DuIE2.0 (49 schemas), and ACE2005 (33 schemas). For the evaluation, we test the model using two strategies: one with the full set of schema queries and another with a fixed set of 10 schema queries. The results depicted in Figure 3 (a) indicate that the mismatch in the number of schema queries during evaluation significantly reduces the model's performance. Further analysis of the model's outputs reveals that the model always tends to generate outputs for each inquiry. We hypothesize that the number of schema queries is one of the key factors affecting the generalization ability. The model needs to first adapt to the number of schema inquiries that are rare during the training and then adapt to the unseen schema.

Inadequate Differentiation Among Schemas Lead to Semantic Similar Confusion. We also evaluate the impact of removing the "Hard Negative Schema Dictionary" on the performance of Baichuan2-IEPILE, with particular attention to schemas that are hard to differentiate. According to the results in Figure 3 (b), we notice that the hard negative schema dictionary plays a relatively limited role in the NER task, which may be due to the clear boundaries inherent to entity recognition. However, the utilization of the hard negative schema dictionary notably enhances model performance in the DuIE2.0 and DuEE1.0 datasets. We observe that semantically similar and easily confused schemas frequently appeared in the model's outputs, such as predicting "dismissal" and "resignation" in the event of "layoff". Therefore, processing instructions that are semantically prone to confusion poses significant challenges, and the hard negative schema dictionary plays a crucial role in bolstering model robustness and improving the accuracy of predictions.

4 Conclusion and Future Work

In this paper, we introduce IEPILE, by collecting and cleaning existing Information Extraction (IE) datasets and utilizing a schema-based instruction generation strategy. Experimental results indicate that IEPILE can help enhance the zero-shot generalization capabilities of LLMs in instruction-based IE. In the future, we will continue to maintain the corpus and try to integrate new resources including open-domain IE, and document-level IE.

Limitations

From the data perspective, our study primarily focuses on schema-based IE, which limits our ability to generalize to human instructions that do not follow our specific format requirements. Additionally, our work is limited to two languages and does not address Open Information Extraction(Open IE), though we plan to extend to more languages and Open IE scenarios in the future. From the model's perspective, our research evaluates limited models, along with a few baselines due to the computation resources. Theoretically, IEPILE can be applied to any other LLMs, such as ChatGLM (Du et al., 2022) and Gemma (Mesnard et al., 2024).

Ethical Considerations

In this paper, we strictly adhered to the standards and principles of ethics. All data collected are from publicly available materials, ensuring the transparency and legality of the research. We thoroughly review the data, verifying the legitimacy of their sources and compliance with their usage, thus avoiding any infringement on personal privacy or involvement with unauthorized information.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their thoughtful and constructive feedback. This work was supported by the National Natural Science Foundation of China (No. 62206246, No. NSFCU23B2055, No. NSFCU19B2027), the Fundamental Research Funds for the Central Universities (226-2023-00138), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Yongjiang Talent Introduction Programme (2021A-156-G), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. This work was supported by Ant Group and Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph.

References

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. Learning to rank context for named entity recognition using a synthetic dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10372–10382. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *NeurIPS 2020*.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning, CoNLL 2004, Held in cooperation with HLT-NAACL 2004, Boston, Massachusetts, USA, May 6-7, 2004*, pages 89–97. ACL.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3470–3479. Association for Computational Linguistics.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. 2022a. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3329–3339. Association for Computational Linguistics.

Xiang Chen, Lei Li, Yuqi Zhu, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, Ningyu Zhang, and Huajun Chen. 2024. Sequence labeling as non-autoregressive dual-query set generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 45–57. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. 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 5882–5888. Association for Computational Linguistics.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics*, 47:1–10.

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

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

Zhaoye Fei, Yunfan Shao, Linyang Li, Zhiyuan Zeng, Hang Yan, Xipeng Qiu, and Dahua Lin. 2024. Query of CC: unearthing large scale domain-specific knowledge from public corpora. *CoRR*, abs/2401.14624.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction. *CoRR*, abs/2310.05092.

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

Tongfeng Guan, Hongying Zan, Xiabing Zhou, Hongfei Xu, and Kunli Zhang. 2020. Cmeie: Construction and evaluation of chinese medical information extraction dataset. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 270–282. Springer.

Honghao Gui, Shuofei Qiao, Jintian Zhang, Hongbin Ye, Mengshu Sun, Lei Liang, Huajun Chen, and Ningyu Zhang. 2023. Instructie: A bilingual instruction-based information extraction dataset. *CoRR*, abs/2305.11527.

Harsha Gurulingappa, Abdul Mateen Rajput, and Luca Toldo. 2012. Extraction of adverse drug effects from medical case reports. *J. Biomed. Semant.*, 3:15.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 172–183. Springer.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4803–4809. Association for Computational Linguistics.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David A. Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 33–38. The Association for Computer Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2023. A reevaluation of event extraction: Past, present, and future challenges. *CoRR*, abs/2311.09562.

Wenhao Huang, Qianyu He, Zhixu Li, Jiaqing Liang, and Yanghua Xiao. 2024. Is there a one-model-fits-all approach to information extraction? revisiting task definition biases.

Sharmistha Jat, Siddhesh Khandelwal, and Partha P. Talukdar. 2017. Improving distantly supervised relation extraction using word and entity based attention. In *6th Workshop on Automated Knowledge Base Construction, AKBC@NIPS 2017, Long Beach, California, USA, December 8, 2017*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 10030–10051. 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.

Veysel Kocaman and David Talby. 2020. Biomedical named entity recognition at scale. In *Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part I*, volume 12661 of *Lecture Notes in Computer Science*, pages 635–646. Springer.

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

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022a. Deduplicating training data makes language models better. 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 8424–8445. Association for Computational Linguistics.

Meisin Lee, Lay-Ki Soon, Eu-Gene Siew, and Ly Fie Sugianto. 2022b. Crudeoilnews: An annotated crude oil news corpus for event extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 465–479. European Language Resources Association.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22-23, 2006*, pages 108–117. Association for Computational Linguistics.

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

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023b. Codeie: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15339–15353. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 894–908. Association for Computational Linguistics.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019. Duie: A large-scale chinese dataset for information extraction. In *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, volume 11839 of *Lecture Notes in Computer Science*, pages 791–800. Springer.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020a. A unified MRC framework for named entity recognition. In *ACL 2020*, pages 5849–5859. Association for Computational Linguistics.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020b. Duee: A large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part II*, volume 12431 of *Lecture Notes in Computer Science*, pages 534–545. Springer.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and James R. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8386–8390. IEEE.

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

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.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. PIVOINE:

instruction tuning for open-world information extraction. *CoRR*, abs/2305.14898.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022a. 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.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022b. Unified structure generation for universal information extraction. In *ACL 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 3219–3232. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10572–10601. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *ICLR 2021*. OpenReview.net.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 548–554. The Association for Computational Linguistics.

Sameer S. Pradhan and Nianwen Xue. 2009. Ontonotes: The 90% solution. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA, Tutorial Abstracts*, pages 11–12. The Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinform.*, 30(6):868–875.

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.

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

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.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron C. Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5571–5587. Association for Computational Linguistics.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational

Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 7072–7079. AAAI Press.

Simone Tedeschi and Roberto Navigli. 2022. Multinerd: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 801–812. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2023. Prompting palm for translation: Assessing strategies and performance. 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 15406–15427. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

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

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: in-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3534–3547. Association for Computational Linguistics.

Haitao Wang, Zhengqiu He, Jin Ma, Wenliang Chen, and Min Zhang. 2019. Ipre: a dataset for interpersonal relationship extraction. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 103–115. Springer.

Jiaqi Wang, Yuying Chang, Zhong Li, Ning An, Qi Ma, Lei Hei, Haibo Luo, Yifei Lu, and Feiliang Ren. 2024. Techgpt-2.0: A large language model project to solve the task of knowledge graph construction.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. GPT-NER: named entity recognition via large language models. *CoRR*, abs/2304.10428.

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.

Zengzhi Wang, Rui Xia, and Pengfei Liu. 2023c. Generative AI for math: Part I - mathpile: A billion-token-scale pretraining corpus for math. *CoRR*, abs/2312.17120.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, pages 2247–2256. IEEE.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023. YAYI-UIE: A chat-enhanced instruction tuning framework for universal information extraction. *CoRR*, abs/2312.15548.

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

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *CoRR*, abs/2312.17617.

Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. CLUENER2020: fine-grained named entity recognition dataset and benchmark for chinese. *CoRR*, abs/2001.04351.

Aiyuan Yang, Bin Xiao, Bingning Wang, et al. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R. Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, Heng Ji, and Chengxiang Zhai. 2024. If LLM is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *CoRR*, abs/2401.00812.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *CoRR*, abs/1508.01006.

Jiasheng Zhang, Xikai Liu, Xinyi Lai, Yan Gao, Shusen Wang, Yao Hu, and Yiqing Lin. 2023a. 2iner: Instructive and in-context learning on few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3940–3951. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023b. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1554–1564. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *ACL 2017*, pages 1227–1236. Association for Computational Linguistics.

Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li. 2021. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. 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*, pages 14638–14646. AAAI Press.

A Related Work

A.1 Information Extraction Datasets

Large-scale pre-trained corpora are crucial for the effectiveness of LLMs, providing a wealth of knowledge and a foundation for language comprehension. At the same time, the annotated data for information extraction (IE) also holds its importance. Although the field of IE has accumulated a considerable amount of annotated data (Walker et al., 2006; Riedel et al., 2010; Sang and Meulder, 2003; Luan et al., 2018; Gui et al., 2023), these datasets are often limited in size, scattered in distribution, and lack standardization in schema. Faced with these limitations, there is an urgent need for generating instruction data through unified and automated methods to bridge the gap presented by the current absence of centralized, large-scale IE instruction datasets. In this paper, we concentrate on instruction-based IE scenarios. We develop a comprehensive, schema-rich instruction dataset for IE by collecting and cleaning existing IE datasets, called IEPILE. IEPILE is designed to enhance the adaptability and processing capabilities of LLMs for different IE tasks, simultaneously strengthening their generalization skills to extract from new domains and schemas.

A.2 Information Extraction Models

Recently, LLMs (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023a,b) demonstrate their exceptional versatility and generalization capabilities across a variety of downstream tasks (Vilar et al., 2023; Hegselmann et al., 2023). Particularly in the domain of IE, these models have the potential to tackle many challenges previously encountered in research (Zheng et al., 2017; Li et al., 2020a; Paolini et al., 2021; Lu et al., 2022b; Lou et al., 2023; Chen et al., 2022b, 2024), such as adaptability issues when dealing with unseen labels. Some studies (Wei et al., 2023; Wang et al., 2023a; Xie et al., 2023) make significant performance gains in low-resource settings by designing prompt-based frameworks and leveraging models like ChatGPT for in-context learning. Moreover, research efforts such as InstructUIE (Wang et al., 2023b), PIVOINE (Lu et al., 2023), and YAYI-UIE (Xiao et al., 2023), which employ instruction-tuning of open-source LLMs, also achieve notable successes on IE. Additional research explore areas such as prompt learning (Zhang et al., 2023a), guidelines (Sainz et al., 2023) and synthetic dataset (Amalvy et al., 2023). Despite these advancements, current models fine-tuned with instruction data face a major challenge: the coarse schema handling strategies in constructing instructions could potentially impair the models' capacity for generalization.

B Construction Details of IEPILE

B.1 Data Collection and Clean

Data Collection To comprehensively cover various domains and meet the practical demands of information extraction (IE), we collect IE datasets from multiple sources. IEPILE dataset mainly involves bilingual data (Chinese and English) and three IE tasks: Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE). The English part mainly comes from the benchmark dataset IEINSTRUCTIONS (Wang et al., 2023b), while the Chinese data is similar to the Chinese datasets mentioned in the YAYI-UIE (Xiao et al., 2023). It should be noted that our Chinese dataset collection is conducted concurrently with the aforementioned research.

Specifically, the NER datasets include fifteen English datasets such as ACE2005 (Walker et al., 2006), AnatEM (Pyysalo and Ananiadou, 2014), BC2GM (Kocaman and Talby, 2020), BC4CHEMD (Kocaman and Talby, 2020), BC5CDR (Zhang et al., 2023b), CoNLL2003 (Sang and Meulder, 2003), FabNER (Kumar and Starly, 2022), FindVehicle (Guan et al., 2023), GENIA-Ent (Kim et al., 2003), HarveyNER (Chen et al., 2022a), MIT Movie (Liu et al., 2013), MIT Restaurant (Liu et al., 2013), MultiNERD (Tedeschi and Navigli, 2022), NCBI-Disease (Dogan et al., 2014), Ontonotes (Pradhan and Xue, 2009), and three Chinese datasets including MSRA (Levow, 2006), Resume NER (Zhang and Yang, 2018), CLUE NER (Xu et al., 2020). The RE task encompasses eight English datasets including ADE Corpus (Gurulingappa et al., 2012), CoNLL2004 (Carreras and Màrquez, 2004), GIDS (Jat et al., 2017), KBP37 (Zhang and Wang, 2015), NYT (Riedel et al., 2010), NYT11-HRL (Takanobu et al., 2019), SciERC (Luan et al., 2018), Semeval-RE (Hendrickx et al., 2010), and two Chinese datasets, CMeIE (Luan et al., 2018), DuIE2.0 (Hendrickx et al., 2010). The EE task covers three English datasets: ACE2005 (Walker et al., 2006), CASIE (Satyapanich et al., 2020), PHEE (Sun et al., 2022), and two Chinese datasets, DuEE1.0 (Satyapanich et al., 2020), DuEE-fin (Sun et al., 2022). These datasets span various domains such as general, medical, financial, and more. For more detailed statistical information, please refer to Tables 9, 10 and 11.

Data Cleaning During the data cleaning process, we address each dataset individually. Firstly, we calculate the text overlap within each dataset's training, validation, and test sets. If a text is discovered to have multiple occurrences within the same file accompanied by inconsistent annotations, we exclude all corresponding instances from the dataset. Secondly, we compare the text overlap between training, validation, and test sets. If texts from the test set appear previously in the training or validation sets, we would exclude these instances from the training and validation sets. Furthermore, we formulate three heuristic rules to eliminate low-quality and meaningless data:

1) Non-alphabetic characters comprising more than 80% of the text;

2) Text length under five characters without any labels;

3) A high prevalence of stopwords such as 'the,' 'to,' 'of,' etc., exceeding 80%.

We believe that the aforementioned cleaning measures will positively affect model training and enhance its performance. Moreover, our efforts unify data formats across various tasks and conduct a thorough audit of each dataset, creating detailed **data records** that include the volume of data, domains, schemas, and other information. Figure 4 is an example of a data record for Ontonotes.

B.2 Schema-Based Instruction Generation

Hard Negative Schema Construction. As illustrated in Figure 1, assume that dataset $\mathcal{D}$ possesses a predefined label set L. For a given text S, the schemas present in its annotation constitute the positive schema set Pos_L, while others form the negative schema set Neg_L. Inspired by the theory of contrastive learning, we construct a hard negative schema dictionary $\mathcal{K}$, where each key represents a unique schema and the associated value is a collection of schemas that are semantically similar to the key schema. Consequently, the set of hard negative schema, $Hard\_L$, is defined as $\mathcal{K}[Pos\_L]$. However, if Neg_L is composed solely of $Hard\_L$, it would lack a sufficient number of negative instances for the model to learn effectively. Therefore, we define another set of negative schemas, $Other\_L = L - Hard\_L - Pos\_L$. Ultimately, the Neg_L is composed of $Hard\_L$ and a small number of $Other\_L$ (roughly $split\_num$). The

Data Record for Ontonotes

Domain = General
#Schema = 18
Schema = [date, organization, person, geographical social political,...]
#Instance = {"train": 54994, "dev": 7997, "test": 7782}
Split_num = 6
#Instruction = 98233
Split_distribution = {"3": 4735, "4": 1759, "5": 7725, "6": 74437, "7": 5736, "8" :3841}

Figure 4: An exemplar of data records for OntoNotes: the domain, the number and details of schemas, the total volume of data, the $split\_num$, the number of instructions produced using our method, along with the distribution of split count within the interval [($split\_num$ / 2), ($split\_num$ + $split\_num$ / 2)].

rationale behind the development of these hard negatives is two-fold: firstly, to induce a more frequent co-occur of semantically similar schemas within the instructions, and secondly, to reduce the volume of training instances without sacrificing the model's performance. In the context of a dataset comprising 48 schemas with a given $split\_num$ of 4, traditional mode would dictate the creation of 12 unique instructions per data point. However, through the integration of hard negatives, this requisite can be substantially minimized to a mere 3 instructions.

Batched Instruction Generation. Subsequently, we obtain the final schema set $L' = Pos\_L + Neg\_L$. During the instruction generation phase, the role of schemas is critically vital, as it reflects the specific extraction requirements and is dynamically variable. Traditional practices typically integrate the full schema set into the instruction. However, in this study, we employ a batched instruction generation method, dynamically limiting the number of schemas inquired in each instruction to the number of $split\_num$, which ranges between 4 to 6. Therefore, L' will be divided into $|L'|/split\_num$ batches for querying, with each batch querying $split\_num$ schemas. Consequently, even if the number of schemas inquired during the evaluation phase differs from that of training, the batched mechanism allows us to distribute the inquiries across $split\_num$ schemas, thereby mitigating the decline in generalization performance.

Selection of split_num. In the determination of the optimal range for $split\_num$, our methodology integrates empirical results with an in-depth analysis of dataset characteristics. For a dataset containing N different labels, the theoretical value of $split\_num$ should fall within the interval [1, N]. Addressing datasets with heterogeneous label counts, our objective is to identify a $split\_num$ value that offers broad applicability across numerous datasets, thus ensuring this value serves as a common divisor for the majority of dataset label counts. For instance, for Named Entity Recognition datasets, we set $split\_num$ to 6; for Relation Extraction and Event Extraction datasets, we establish $split\_num$ at 4. We also observe that when $split\_num$ is 1, the ratio of positive to negative samples significantly impacts model performance, and the corresponding number of training samples becomes vast, affecting efficiency adversely. More crucially, we believe that enumerating multiple schemas in instructions aids the model in more effectively learning to distinguish and identify various schemas, thereby enhancing model performance.

Furthermore, to enhance model robustness and its clear understanding of the dynamically changing schema sequences in instructions, we set the actual number of schema splits within a dynamic range of [$split\_num$ // 2, $split\_num$ + $split\_num$ // 2]. Specifically, if the number of schemas in the last batch is less than half of $split\_num$, it is merged with the previous batch; otherwise, it stands as an independent batch.

Instruction Format The instruction format of IEPILE adopts a structure akin to JSON strings, essentially constituting a dictionary-type string. This structure is comprised of three main components: (1) "instruction", which is the task description outlining the objective of the instruction's execution; (2) "schema", a list of labels that need to be extracted; (3) "input", the source text from which information is to be extracted. Examples of instructions corresponding to various tasks can be found in Table 12.

B.3 Data Statistics

Based on the aforementioned methodologies, we construct a high-quality information extraction instruction dataset known as IEPILE. This dataset contains approximately two million instances and approximately 0.32B tokens. Each instance of IEPILE comprises two fields: "instruction" and "output", which are formatted for direct use in the instruction tuning.

C Experiments

C.1 Experimental Settings

Evaluation Metrics We employ span-based Micro-F1 as the primary metric for measuring model performance. For the NER task, the model is required to accurately identify the boundaries of entities and their corresponding types. For the RE task, the model must precisely determine the subject and object entities within a relation, as well as the type of relation between them. UIE necessitates predefined entity types; given that the FewRel and Wiki-ZSL datasets do not provide such information, we are unable to evaluate UIE's performance on these datasets. As for the EE task, we match the event triggers, denoted as **Trigger**, and the arguments, referred to as **Argument**, independently.

Baseline models To assess the zero-shot generalization capabilities, we select a range of strong models for comparative analysis:

- UIE (Lu et al., 2022a): is a unified text-to-structure generation framework that can model various information extraction (IE) tasks generically.
- LLaMA2-13B-Chat (Touvron et al., 2023b): is a series of LLMs ranging from 7 billion to 70 billion parameters.
- Baichuan2-13B-Chat (Yang et al., 2023): is a collection of multilingual LLMs containing 7 billion and 13 billion parameters.
- Qwen1.5-14B-Chat (Bai et al., 2023): is a comprehensive language model series that encompasses distinct models with varying parameter counts.
- Mistral-7B-Instruct-v0.2 (Jiang et al., 2023): is a 7-billion-parameter LLM.
- ChatGPT (Ouyang et al., 2022): also known as GPT-3.5-turbo, represents the most advanced artificial intelligence language model with chat optimization capabilities to date.
- GPT-4 (OpenAI, 2023): Known as the most powerful closed-source chat model to date.
- LLaMA3-8B-Instruct [3]: The latest release in the LLaMA model series, achieving significant improvements across various benchmarks.
- InstructUIE (Wang et al., 2023b): a unified IE framework based on multi-task instruction tuning.
- YAYI-UIE (Xiao et al., 2023): is an end-to-end, chat-enhanced, universal information extraction framework that supports both Chinese and English, fine-tuned with instructional prompts for generalized information.

C.2 OneKE

We leverage IEPILE, InstructIE (Gui et al., 2023), CMRC (Cui et al., 2019), along with certain proprietary business information extraction datasets from Ant Group, to compile a comprehensive training dataset consisting of 2.5 million instances. Subsequently, we undertake full-parameter fine-tuning of the alpaca2-chinese-13b[4] model on this training dataset, resulting in the refined model named OneKE.

Zero-shot Dataset To ensure the validity of the zero-shot evaluation and prevent result bias due to data similarity, we select datasets primarily derived from news and biomedical fields as our training sets. This selection is intended to train the model's capability for instruction following and schema-based extraction. For the evaluation data, we adopt the 13 cross-domain datasets recommended in IE-INSTRUCTIONS and YAYI-UIE, which include: for Named Entity Recognition (NER) tasks, we use the CrossNER (Liu et al., 2021), Weibo NER (Peng and Dredze, 2015), and Boson[5]; in Relation Extraction (RE) tasks, we choose FewRel (Han et al., 2018), Wiki-ZSL (Chen and Li, 2021), COAE2016[6], IPRE (Wang et al., 2019), and SKE2020[7]; and for Event Extraction (EE), we include RAMS (Ebner et al., 2020), WikiEvents (Li et al., 2021), CrudeOilNews (Lee et al., 2022b), FewFC (Zhou et al., 2021), and CCF law [8]. These

[3] https://ai.meta.com/blog/meta-llama-3/.
[4] https://huggingface.co/hfl/chinese-alpaca-2-13b.
[5] https://github.com/InsaneLife/
[6] https://github.com/Sewens/COAE2016
[7] https://aistudio.baidu.com/datasetdetail/177191
[8] https://aistudio.baidu.com/projectdetail/4201483

	Method	EN				CH		
		WikiEvents	RAMS	CrudeOil News	Avg	FewFC	CCF Law	Avg
Trigger	LLaMA2	0.00	0.00	0.00	0.00	0.23	0.08	0.16
	Baichuan2	0.00	0.00	0.00	0.00	11.82	2.73	7.28
	Qwen1.5	0.00	0.00	0.00	0.00	11.47	3.25	7.36
	Mistral	0.00	0.00	0.00	0.00	4.69	0.23	2.46
	ChatGPT	2.95	8.35	1.41	4.24	16.15	0.00	8.08
	GPT4.0	5.24	10.14	26.13	13.84	74.25	42.12	58.19
	UIE	5.12	9.25	6.45	6.94	-	-	-
	InstructUIE	11.64	**24.27**	23.26	19.72	-	-	-
	YAYI-UIE	10.97	18.87	12.45	14.10	81.28	12.87	47.08
	Baichuan2-IEPILE	9.12	20.19	36.61	21.97	**83.59**	**63.53**	**73.56**
	LLaMA2-IEPILE	**13.93**	23.62	33.87	23.81	70.10	59.90	65.00
	Qwen1.5-IEPILE	11.38	21.26	30.69	21.11	78.77	61.43	70.10
	LLaMA3-IEPILE	9.71	20.27	**39.88**	23.29	81.52	59.92	70.72
	OneKE	12.43	22.58	38.49	**24.50**	80.11	62.19	71.15
Argument	LLaMA2	0.00	0.00	0.00	0.00	0.00	0.06	0.03
	Baichuan2	0.79	1.81	0.48	1.03	6.91	13.04	9.98
	Qwen1.5	0.64	2.31	0.74	1.23	6.37	14.48	10.43
	Mistral	0.24	0.65	0.16	0.35	7.43	6.60	7.02
	ChatGPT	2.07	2.21	8.60	4.29	44.40	44.57	44.49
	GPT4.0	3.35	7.35	17.25	9.32	48.05	47.49	47.77
	UIE	1.78	2.14	8.95	4.29	-	-	-
	InstructUIE	5.88	6.21	**21.78**	11.29	-	-	-
	YAYI-UIE	5.11	8.21	19.74	11.02	**63.06**	59.42	61.24
	Baichuan2-IEPILE	7.64	10.42	20.40	12.82	57.93	**65.43**	**61.68**
	LLaMA2-IEPILE	**12.55**	11.30	18.47	14.11	43.26	35.71	39.49
	Qwen1.5-IEPILE	11.93	10.57	20.22	14.24	59.49	58.86	59.18
	LLaMA3-IEPILE	12.10	10.96	19.20	14.09	48.19	42.59	45.39
	OneKE	11.88	**13.26**	20.11	**15.08**	58.83	62.38	60.61

Table 3: Zero-shot performance on Event Extraction (EE) task. Within each column, shadow and shadow represent the top 2 results.

datasets cover a wide range of fields including literature, music, law, and oil news. It is noteworthy that these evaluation data sets are not used during the training, ensuring that our evaluation accurately reflects the model's generalization and adaptation capabilities for unseen domains and unseen schema data in zero-shot information extraction.

C.3 Zero-shot performance on Event Extraction

As illustrated in Table 3, the model trained with IEPILE exhibits outstanding performance in zero-shot event extraction (EE) tasks, surpassing other baselines. Notably, in the Chinese EE task, the LLaMA2-IEPILE model's performance is slightly inferior to YAYI-UIE's, revealing LLaMA2's limitations in processing Chinese data. However, in the English EE task, LLaMA2-IEPILE's performance is significantly superior to that of similar models. This contrast highlights the potential influence of language type on model performance.

C.4 Hyper-parameter

In our research, we select four pre-trained models, Baichuan2-13B-Chat and LLaMA2-13B-Chat, Qwen1.5-14B-Chat, and LLaMA3-8B-Instruct, as the base models for our study. Specifically, we employ the LoRA (Hu et al., 2022) technique and utilize 8 NVIDIA A800 GPUs to perform instruction tuning on our IEPILE dataset. Detailed configurations of the hyperparameters during the fine-tuning process are presented in Table 4.

Hyperparameter	Value
Number of Epochs	5
Learning Rate	5e-5
Batch Size	20
Accumulate	4
Lora_r	64
Lora_alpha	64
Lora_dropout	0.05

Table 4: Training Hyperparameters

Dataset	**Supervised**	**Zero-shot**
ACE2004	84.28	77.01
People Daily	98.34	95.29

Table 5: The results of individual LoRA fine-tuning on ACE2004 and People Daily datasets for Baichuan2-13B-Chat, compared with the zero-shot generalization results of Baichuan2-IEPILE on these two datasets.

C.5 Supervision Results

Due to limited computational resources, I report only the supervised results for the Baichuan2-IEPILE, LLaMA2-IEPILE, and OneKE models. Tables 6, 7, and 8 present our experimental results under a supervised learning setting on the training dataset. Specifically, it can be observed that after training on the IEPILE, the model excels in Named Entity Recognition (NER), Relation Extraction (RE), and Event Detection (ED), ranking top 2 across these tasks. The model's performance is only slightly behind other baselines in the Event Argument Extraction. Additionally, we record the model's performance in Chinese NER, RE, and EE tasks, where it demonstrates robust results. In a comprehensive assessment, the IEPILE-trained model showcases performance on par with other models in instruction-based information extraction (IE) tasks and significantly improves performance in zero-shot IE tasks compared to other models. This indicates the significant application prospects and potential of IEPILE in the current field of IE.

C.6 Impact of Potential Dataset Bias on Model Performance and Generalization

During the research, we identify that potential biases introduced by the datasets used can affect the model's performance and generalization capability. Firstly, biases in the definition of schemas within the datasets have a negative impact on model performance (Huang et al., 2024). In the early stages of training, we observe instability in results due to mutual interference among multiple datasets that contain the same schemas but with differing definitions. For instance, despite wikiann, wikineural, polyglot-NER, and CoNLL2003 all containing common schemas such as people and organization, they each possess distinct scheme definitions. Consequently, in the later stages, only CoNLL2003 is retained. Secondly, the model demonstrates good generalization when dealing with datasets having schemas similar to those in the training set. As shown in Table 5, despite not being included in the training corpus, the People Daily and ACE2004 NER datasets share similar schemas with the MASR and ACE2005 NER dataset in the training set, and the Baichuan2-IEPILE model is still capable of handling them proficiently. Lastly, the use of common, coarse-grained labels (such as "person" and "organization") within the IEPILE lead the model, after training, to favor these coarse categories over fine-grained ones (such as "scientist" and "company") when predicting instructions that included both levels of granularity.

Dataset	InstructUIE	YAYI-UIE	Baichuan2-IEPILE	LLaMA2-IEPILE	OneKE
ACE2005	**86.66**	81.78	81.86	81.14	83.45
AnatEM	**90.89**	76.54	87.21	86.90	87.88
BC2GM	**85.16**	82.05	80.73	83.07	82.05
BC4CHEMD	90.30	88.46	90.45	90.07	**90.56**
BC5CDR	**89.59**	83.67	88.07	88.01	88.45
CoNLL2003	92.94	**96.77**	92.49	92.98	93.04
FabNER	76.20	72.63	77.07	76.33	**81.06**
FindVehicle	89.47	98.47	98.49	97.91	**99.45**
GENIA-Ent	74.71	75.21	76.66	77.32	**78.29**
HarveyNER	**88.79**	69.57	67.70	62.64	69.87
MIT Movie	89.01	70.14	88.23	89.54	**89.96**
MIT Restaurant	**82.55**	79.38	79.85	81.30	79.89
MultiNERD	92.32	88.42	94.60	94.24	**94.69**
NCBI-Disease	**90.23**	87.29	85.26	87.59	86.95
Ontonotes	90.19	87.04	87.55	**90.34**	89.08
Avg	**87.27**	82.49	85.08	85.29	86.24
MSRA	-	**95.57**	87.99	86.32	89.02
Resume NER	-	-	93.92	92.86	**95.84**
CLUE NER	-	-	**80.19**	76.57	78.43

Table 6: Overall supervision results on Named Entity Recognition (NER) datasets. Within each row, shadow and shadow represent the top 2 results.

Dataset	InstructUIE	YAYI-UIE	Baichuan2-IEPILE	LLaMA2-IEPILE	OneKE
ADE Corpus	82.31	84.14	83.73	85.87	**87.24**
CoNLL2004	78.48	**79.73**	72.87	73.71	76.16
GIDS	**81.98**	72.36	74.71	74.13	76.69
KBP37	36.14	59.35	65.09	61.49	**65.23**
NYT	90.47	89.97	93.00	92.22	**94.04**
NYT11-HRL	56.06	**57.53**	53.19	54.86	55.56
SciERC	45.15	40.94	43.53	44.58	**45.89**
Semeval-RE	**73.23**	61.02	58.47	57.61	61.46
Avg	67.98	68.13	68.07	68.06	**70.28**
CMeIE	-	-	49.16	47.40	**49.54**
DuIE2.0	-	**81.19**	75.61	74.34	75.73

Table 7: Overall supervision results on Relation Extraction (RE) datasets. Within each row, shadow and shadow represent the top 2 results.

	Dataset	InstructUIE	YAYI-UIE	Baichuan2-IEPILE	LLaMA-IEPILE	OneKE
Trigger	ACE2005	**77.13**	65.00	72.46	70.63	71.17
	CASIE	**67.80**	63.00	60.07	61.27	63.82
	PHEE	**70.14**	63.00	66.22	68.52	68.60
	Avg	**71.69**	63.67	66.25	66.81	67.86
	DuEE1.0	-	85.00	**86.73**	84.01	85.75
	DuEE-fin	-	82.50	**83.54**	79.00	82.91
Argument	ACE2005	**72.94**	62.71	63.90	62.69	62.75
	CASIE	63.53	**64.23**	56.07	56.78	57.16
	PHEE	62.91	**77.19**	70.85	71.33	72.84
	Avg	66.46	**68.04**	63.60	63.61	64.25
	DuEE1.0	-	**79.08**	75.63	73.79	75.40
	DuEE-fin	-	70.02	**79.34**	73.08	78.98

Table 8: Overall supervision results on Event Extraction (EE) datasets. Within each row, shadow and shadow represent the top 2 results.

Task	Dataset	Domain	#Schemas	#Train	#Val	#Test
NER-en	AnatEM (Pyysalo and Ananiadou, 2014)	Biomedical	1	5667	2081	3758
	BC2GM (Kocaman and Talby, 2020)	Biomedical	1	12392	2483	4977
	BC4CHEMD (Kocaman and Talby, 2020)	Biomedical	1	30488	30468	26204
	NCBI-Disease (Dogan et al., 2014)	Biomedical	1	5432	923	940
	BC5CDR (Zhang et al., 2023b)	Biomedical	2	4545	4569	4788
	HarveyNER (Chen et al., 2022a)	Social Media	4	3553	1270	1260
	CoNLL2003 (Sang and Meulder, 2003)	News	4	12613	3070	3184
	GENIA (Kim et al., 2003)	Biomedical	5	14966	1657	1850
	ACE2005 (Walker et al., 2006)	News	7	7134	964	1050
	MIT Restaurant (Liu et al., 2013)	Social Media	8	7658	-	1520
	MIT Movie (Liu et al., 2013)	Social Media	12	9707	-	2441
	FabNER (Kumar and Starly, 2022)	Scientific	12	9421	2179	2064
	MultiNERD (Tedeschi and Navigli, 2022)	Wikipedia	16	130623	9994	9994
	Ontonotes (Pradhan and Xue, 2009)	General	18	54994	7997	7782
	FindVehicle (Guan et al., 2023)	Traffic	21	21547	-	20769
	CrossNER_Politics† (Liu et al., 2021)	Political	9	-	-	650
	CrossNER_Literature† (Liu et al., 2021)	Literary	12	-	-	416
	CrossNER_Music† (Liu et al., 2021)	Musical	13	-	-	465
	CrossNER_AI† (Liu et al., 2021)	AI	14	-	-	431
	CrossNER_Science† (Liu et al., 2021)	Scientific	17	-	-	543
NER-zh	MSRA NER (Levow, 2006)	News	3	40500	4500	3437
	Resume NER (Zhang and Yang, 2018)	Resume	8	3799	463	476
	CLUE NER (Xu et al., 2020)	News	10	9674	1074	1343
	Weibo NER† (Peng and Dredze, 2015)	News	4	-	-	258
	Boson† 5	News	6	-	-	191

Table 9: Statistical data of Named Entity Recognition (NER) datasets, with an † indicating the zero-shot evaluation set not included in the training. CrossNER (Liu et al., 2021) is divided into five subsets for our statistical analysis.

Task	Dataset	Domain	#Schemas	#Train	#Val	#Test
RE-en	ADE Corpus (Gurulingappa et al., 2012)	Biomedical	1	3416	427	428
	GIDS (Jat et al., 2017)	News	4	8525	1417	4307
	CoNLL2004 (Carreras and Màrquez, 2004)	News	5	922	231	288
	SciERC (Luan et al., 2018)	Scientific	7	1366	187	397
	Semeval-RE (Hendrickx et al., 2010)	Scientific	10	6478	1492	2714
	NYT11-HRL (Takanobu et al., 2019)	News	12	60765	146	362
	KBP37 (Zhang and Wang, 2015)	News	18	15911	1723	3405
	NYT (Riedel et al., 2010)	News	24	54412	4975	4985
	Wiki-ZSL (Chen and Li, 2021) †	Wikipedia	83	-	-	-
	FewRel (Han et al., 2018) †	Wikipedia	100	-	-	-
RE-zh	CMeIE (Guan et al., 2020)	Biomedical	53	14339	3585	-
	DuIE2.0 (Li et al., 2019)	News	49	171126	20652	-
	COAE2016† 6	General	9	-	-	971
	IPRE† (Wang et al., 2019)	General	35	-	-	3340
	SKE2020† 7	News	49	-	-	3601

Table 10: Statistical data of Relation Extraction (RE) datasets, with an † indicating the zero-shot evaluation set not included in the training. The test sets for CMeIE and DuIE2.0 are not open-sourced, thus we use the validation sets as our evaluation set. For the FewRel and Wiki-ZSL datasets, we follow Chia et al. (2022).

Task	Dataset	Domain	#Schemas	#Train	#Val	#Test
EE-en	ACE2005 (Walker et al., 2006)	News	33(22)	3257	319	293
	CASIE (Satyapanich et al., 2020)	Cybersecurity	5(26)	3732	777	1492
	PHEE (Sun et al., 2022)	Biomedical	2(16)	2897	960	968
	CrudeOilNews † (Lee et al., 2022b)	Oil News	18(104)	-	-	356
	RAMS † (Ebner et al., 2020)	News	106(398)	-	-	887
	WikiEvents † (Li et al., 2021)	Wikipedia	31(81)	-	-	249
EE-zh	DuEE1.0 (Li et al., 2020b)	News	65(217)	11908	1492	-
	DuEE-Fin (Han et al., 2022)	Finance	13(91)	7015	1171	-
	FewFC † (Zhou et al., 2021)	Finance	5(29)	-	-	2879
	CCF law †8	Law	9(39)	-	-	971

Table 11: Statistical data of Event Extraction (EE) datasets, with an † indicating the zero-shot evaluation set not included in the training. The test sets for DuEE1.0 and DuEE-Fin are not open-sourced, thus we use the validation sets as our evaluation set.

Task	Instruction & Output

NER

```
{
    "instruction": "You are an expert in named entity recognition. Please extract entities that match the schema definition from the input. Return an empty list if the entity type does not exist. Please respond in the format of a JSON string.",
    "schema": ["location", "else", "organization", "person"],
    "input": "The objective of the Basic Course on War is to provide for combatants of the EPR basic military knowledge for the armed conflict against the police and military apparatus of the bourgeoisie."
}
output = {
    "location": [],
    "else": [],
    "organization": ["EPR"],
    "person": []
}
```

RE

```
{
    "instruction": "You are an expert in relationship extraction. Please extract relationship triples that match the schema definition from the input. Return an empty list for relationships that do not exist. Please respond in the format of a JSON string.",
    "schema": ["place of birth", "country capital", "country of administrative divisions", "company"],
    "input": "Born on May 1 , 1927 , in Brichevo , Bessarabia in the present-day Republic of Moldova , Mr. Bertini emigrated to Palestine with his family as a child and pursued musical studies there , in Milan , and in Paris , where he worked with Nadia Boulanger and Arthur Honegger."
}
output = {
    "place of birth": [{"head": "Mr. Bertini", "tail": "Paris"}],
    "country capital": [],
    "country of administrative divisions": [],
    "company": []
}
```

EE

```
{
    "instruction": "You are an expert in event extraction. Please extract events from the input that conform to the schema definition. Return an empty list for events that do not exist, and return NAN for arguments that do not exist. If an argument has multiple values, please return a list. Respond in the format of a JSON string.",
    "schema": [{"event_type": "pardon", "trigger": true, "arguments": ["defendant"]}, {"event_type": "extradite", "trigger": true, "arguments": ["person", "agent", "destination", "origin"]}, {"event_type": "sue", "trigger": true, "arguments": ["place", "plaintiff"]}, {"event_type": "start position", "trigger": true, "arguments": ["person", "entity", "place"]}],
    "input": "Ethical and legal issues in hiring Marinello"
}
output = {
    "pardon": [],
    "extradite": [],
    "sue": [],
    "start position": [{"trigger": "hiring", "arguments": {"person": "Marinello", "entity": "NAN", "place": "NAN"}}]
}
```

Table 12: Instructions and outputs for 3 tasks: Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE). The instruction and output formats for IEPILE adopt a structure similar to JSON strings.