Retrieval Augmented Instruction Tuning for Open NER with Large Language Models

Tingyu Xie[1*], **Jian Zhang**[1*], **Yan Zhang**[2*], **Yuanyuan Liang**[3], **Qi Li**[1], **Hongwei Wang**[1†]
[1]Zhejiang University, [2]National University of Singapore, [3]East China Normal University
{tingyuxie, jianzhang.22, hongweiwang}@zju.edu.cn, yanzhang.jlu@gmail.com

Abstract

The strong capability of large language models (LLMs) has been applied to information extraction (IE) through either retrieval augmented prompting or instruction tuning (IT). However, the best way to incorporate information with LLMs for IE remains an open question. In this paper, we explore **R**etrieval **A**ugmented **I**nstruction **T**uning (**RA-IT**) for IE, focusing on the task of open named entity recognition (NER). Specifically, for each training sample, we retrieve semantically similar examples from the training dataset as the context and prepend them to the input of the original instruction. To evaluate our RA-IT approach more thoroughly, we construct a Chinese IT dataset for open NER and evaluate RA-IT in both English and Chinese scenarios. Experimental results verify the effectiveness of RA-IT across various data sizes and in both English and Chinese scenarios. We also conduct thorough studies to explore the impacts of various retrieval strategies in the proposed RA-IT framework.[1]

1 Introduction

The powerful generalizability of large language models (LLMs) (OpenAI, 2024; Touvron et al., 2023; Bai et al., 2023) has been widely applied to information extraction (IE) (Sainz et al., 2024; Wang et al., 2023b). The major two lines of works for generative IE with LLMs, are prompt designing with retrieval augmented generation (RAG) using an off-the-shelf LLM (Wang et al., 2023a; Guo et al., 2023; Xie et al., 2024), and task-specific instruction tuning (IT) (Zhou et al., 2024; Sainz et al., 2024; Li et al., 2024). However, the best approach to incorporate information to LLMs for IE remains an open question. Inspired by recent studies on retrieval aware and context-enhanced IT (Jiang et al., 2023; Luo et al., 2023; Zhang et al., 2024; Asai et al., 2024; Lin et al., 2024; Liu et al., 2024) for enhancing the LLM capability in downstream tasks, we conduct an empirical study of exploring **R**etrieval **A**ugmented **IT** (**RA-IT**) for IE, with a focus on the of open NER task.

The previous work UniNER (Zhou et al., 2024) distills the strong capability of ChatGPT in open NER into smaller models through IT without any human-annotated data. We follow this line and investigate RA-IT under this *targeted distillation* setting. Other works of IT for IE like Sainz et al. (2024); Li et al. (2024) using code-style instruction data, are *orthogonal* to this work since RA-IT can be integrated into various instruction styles.

In our RA-IT approach, for each training sample, we retrieve *semantically similar examples* from the training dataset and prepend them to the original instruction, forming the context-enhanced instruction. We also explore the impacts of diverse retrieval strategies. Moreover, we construct a Chinese IT dataset for open NER and evaluate our method in both English and Chinese scenarios. We conduct thorough experiments across various data sizes and obtain the following key findings: (1) RA-IT achieves consistent improvements on various data sizes, suggesting the need for context-enhanced fine-tuning. (2) Retrieving semantically similar examples benefits the most for training among various retrieval strategies. Random retrieval also exhibits improvement but shows inferior performance to similar examples. (3) Retrieving out-domain examples for inference requires applying example filtering strategies to achieve improvements. Providing in-domain examples benefits inference.

Our main contributions are two folds: (1) We empirically study the RA-IT framework for open NER. We prepare the retrieval augmented instruction data with semantically similar examples. We conduct thorough experimental analysis to study the impact of various retrieval strategies. (2) We

[*] Equally contributed.
[†] Corresponding authors.
[1]Code and data are available at: https://github.com/Emma1066/Retrieval-Augmented-IT-OpenNER

Proceedings of the 31st International Conference on Computational Linguistics, pages 2904–2918
January 19–24, 2025.

construct an IT dataset for Chinese open NER and conduct our investigations in English and Chinese scenarios across various data sizes. Experimental results verify the benefits of RA-IT for open NER.

2 Method

Preliminary: Targeted Distillation. We follow UniNER (Zhou et al., 2024) to conduct our study in the setting of targeted distillation, where they successfully distill the strong capability of ChatGPT in open NER into smaller models, without any human-annotated data. The pipeline is as follows: (1) Data construction. They sample inputs from a large corpus across diverse domains, then use ChatGPT to automatically generate NER outputs. (2) Distillation. After obtaining the automatically constructed data, they apply IT to distill the open NER capability of ChatGPT into smaller models.

Vanilla IT. The original instruction tuning template used in targeted distillation is shown in the bottom part of Fig. 1, which we refer to as Vanilla IT, where each passage and its associated entity output are converted into a multi-turn conversation.

RA-IT. We explore an alternative way to conduct IT in targeted distillation: we introduce RA-IT, a context-enhanced tuning approach, of which the overview is in Fig. 1. In our RA-IT approach, each data is augmented with a retrieved context, which consists of k *semantically similar examples* retrieved from the training dataset. The retrieved context is prepended to the original conversation, forming the retrieval augmented instruction. By fine tuning LMs in this recipe, we equip the LMs with the ability to generate NER answer with on-demand RAG. This means we could flexibly adapting LMs to different scenarios by determining whether to use RAG during inference based on the specific characteristics of the scenario.

Retriever. We use sentence embedding-based retrieval and adopt cosine similarity as our similarity metric. We retrieve the k nearest neighbors as context. We also investigate various retrieval strategies for both training and inference stages.

3 Experiment

3.1 Experimental Settings

Backbones: We adopt LLaMA-3-8B (Meta, 2024) and Qwen-1.5-7B (Team, 2024) as the backbone models for English and Chinese scenarios respectively. **Training:** For English, we use the training data Pile-NER released by Zhou et al. (2024). For Chinese, we use the training data Sky-NER constructed in this paper as described in Section 3.2. We use LoRA (Hu et al., 2021) to train models. Our training infrastructure was 1 NVIDIA A100 80GB. **Retrieval:** We adopt GTE-large[2] (Li et al., 2023) to generate text embeddings and set $k = 2$ in main experiments. **Evaluation:** We mainly focus on the zero-shot evaluation. For English, we adopt benchmarks CrossNER, MIT-Movie and MIT-restaurant following Zhou et al. (2024). For Chinese, we collect eight benchmarks across diverse domains, of which details are in Appendix C. We report micro-F1 value.

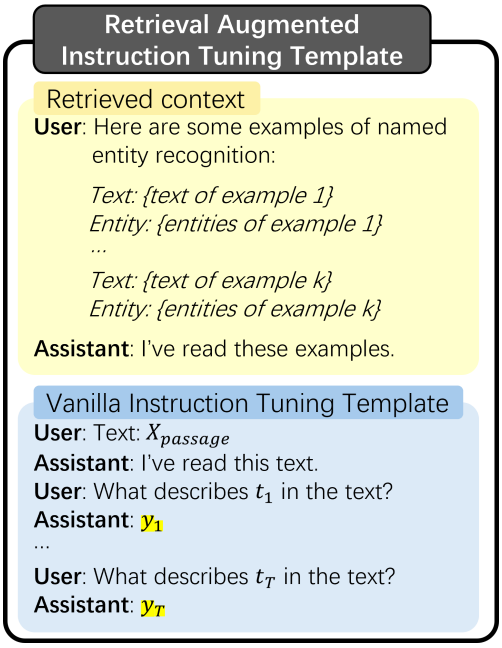

Figure 1: The **RA-IT template**, where the **retrieved context** consists of *semantically similar examples* retrieved from the training dataset and is prepended to the original **vanilla IT template**. The vanilla IT template, presented by Zhou et al. (2024) converts each NER sample into a conversation, where $X_{passage}$ is the input text, $[t1, \ldots, t_T]$ are entity types to extract, and y_i is the list of entity mentions that are t_i. The highlighted parts are used to compute the loss during training.

3.2 Chinese IT Data Construction

Following the data construction recipe of UniNER (Zhou et al., 2024), we construct an IT dataset for Chinese open NER. We sample input passages from the large-scale Sky corpus (Wei et al., 2023) across various domains, then use ChatGPT (gpt-3.5-turbo) to generate entity mentions and types based on the

[2]https://huggingface.co/thenlper/gte-large

Data size	Method	Movie	Restaurant	AI	Literature	Music	Politics	Science	Avg.
-	ChatGPT	5.30	32.80	52.40	39.80	66.60	68.50	67.00	47.50
5K	Vanilla IT	44.87	42.72	52.87	59.00	60.47	59.35	58.36	53.95
	RA-IT	50.26	45.75	52.61	60.01	63.04	60.02	58.91	**55.80**
10K	Vanilla IT	49.81	41.47	53.78	60.99	63.79	60.84	61.47	56.02
	RA-IT	53.79	45.73	55.90	62.58	66.52	62.40	63.67	**58.65**
50K	Vanilla IT	44.83	40.39	58.63	62.88	64.12	61.63	63.22	56.53
	RA-IT	45.18	40.78	58.01	63.60	64.76	61.90	62.79	**56.72**

Table 1: Zero-shot evaluation in English scenario. We report F1 values (%). Numbers in **bold** indicates the best results of each category. RA-IT shows consistent improvements across various dats sizes, suggesting the need of context-enhanced training.

Data size	Method	Ontonotes 4	MSRA	Weibo	Boson	ClueNER	CMeEE	Ren.	Yidu	Avg.
-	ChatGPT	29.70	41.36	30.25	46.65	44.75	43.16	34.25	34.90	38.13
5K	Vanilla IT	48.88	51.47	38.95	52.47	43.54	41.50	47.51	47.23	46.44
	RA-IT	49.23	53.08	37.43	52.64	43.27	43.87	48.31	48.47	**47.04**
10K	Vanilla IT	46.28	52.56	39.26	52.92	45.42	42.59	47.99	47.95	46.87
	RA-IT	47.69	55.06	37.38	53.86	45.25	43.71	49.25	47.86	**47.51**
50K	Vanilla IT	43.99	50.02	34.55	54.98	43.59	42.52	49.37	49.63	46.08
	RA-IT	46.72	54.15	33.28	54.43	43.86	43.78	49.50	50.24	**47.00**

Table 2: Zero-shot evaluation in Chinese scenario. We report F1 values (%). Numbers in **bold** indicates the best results of each category. RA-IT shows consistent improvements across diverse data sizes in Chinese scenario, which further verifies the benefits of our RA-IT approach.

Frequency	Entity Type
Top 1% (75.3%)	概念(concept), 地点(location), 人物(person), 组织(organization), 产品(product)...
1%-10% (17.5%)	荣誉(honor), 技术类(technical), 场所(place), 情绪(emotion), 节目(program)...
10%-100% (7.2%)	比赛组别(competition category), 房产类型(property type)...

Table 3: Statistics of Sky-NER, the constructed IT dataset for Chinese open NER. Example entity types from various frequency ranges - top 1%, 1-10% and 10-100%, along with the percentage of total frequencies for each range.

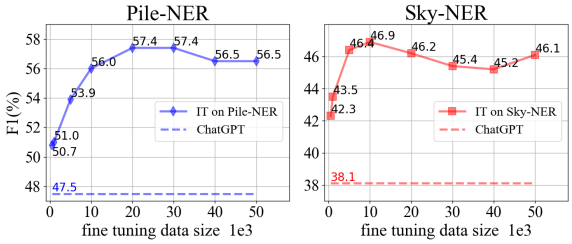

Figure 2: Preliminary study of IT data efficiency for open NER in English (left) and Chinese (right) scenarios, where the training data are Pile-NER and Sky-NER respectively. Average zero-shot results of evaluated benchmarks are illustrated. The performance does not necessarily improve as the data increases.

sampled passages. More details of data construction procedures are in Appendix A. We name this dataset as Sky-NER, which consists of 50K NER examples, and the type statistics are in Table 3.

3.3 Preliminary Study on Data Efficiency

We conduct a preliminary study on IT data efficiency in targeted distillation for open NER by exploring the impact of varous datas sizes: [0.5K, 1K, 5K, 10K, 20K, 30K, 40K, 50K]. We use vanilla IT for preliminary study. Results are visualized in Fig. 2. The following observations are consistent in English and Chinese: (1) a small data size already surpass ChatGPT's performances. (2) Performances are improving as the data sizes increased to 10K or 20K, but begin to decline and then remain at a certain level as data sizes further increased to 50K. Recent work for IT data selection, Xia et al. (2024); Ge et al. (2024); Du et al. (2023) also find the superior performances of only limited data size. We leave selecting more beneficial IT data for IE as future work. Accordingly, we conduct main experiments on 5K, 10K and 50K data sizes.

3.4 Main results

The main results are summarized in Table 1 and 2 respectively. We report the results of inference without examples for RA-IT here, since we found this setting exhibits more consistent improvements. The impacts of inference with examples are studied in Section 3.5. As shown in the tables, RA-IT

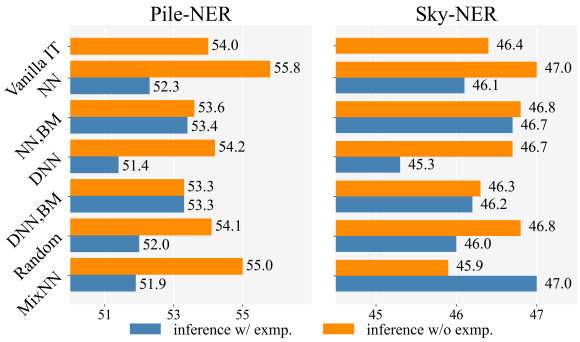

Figure 3: Impacts of training using various retrieval strategies in RA-IT. The average F1 value of the evaluated benchmarks is reported. *NN* exhibits the best performances, suggesting the need of training with retrieved context.

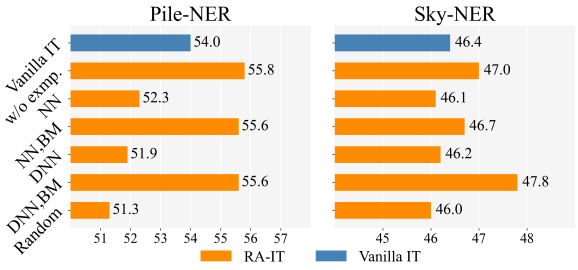

Figure 4: Impacts of inferece with *out-domain examples* using various retrieval strategies. The average F1 value of the evaluated benchmarks are reported. **w/o exmp.** means inference without example. Applying example filtering strategy such as BM25 filtering benefits RAG with out-domain examples.

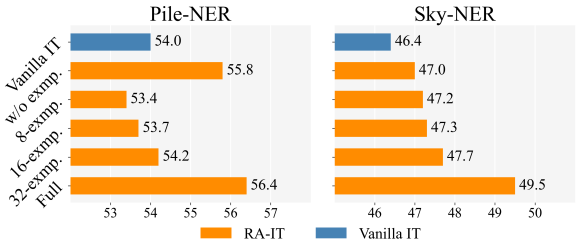

Figure 5: Impacts of inference with *in-domain examples* using *NN* retrieval. The average F1 value of the evaluated benchmarks are reported. N**-exmp.** means the example pool of size N. The results indicate that sufficient in-domain examples are helpful for inference with RAG.

shows consistent improvements on English and Chinese across various data sizes. This presumably because the retrieved context enhance the model ability to understand the inputs. This suggests the need for context-enhanced instructions.

3.5 Analysis

We explore the impacts of diverse retrieval strategies. We conduct analysis on 5K data size for cost saving as the effect of RA-IT is consistent across various data sizes as shown in Section 3.4. We report the average results of the evaluated benchmarks here.

Diverse retrieval strategies. The following strategies are explored in the subsequent analysis. (1) Nearest neighbor (*NN*), the strategy used in the main experiments, retrieves k nearest neighbors of the current sample. (2) Nearest neighbor with BM25 filter (*NN, BM*), where we apply BM25 scoring to filters out NN examples not passing a predefined threshold. Samples with no satisfied examples are used with the vanilla instruction template. (3) Diverse nearest neighbor (*DNN*), retrieves K nearest neighbors with $K >> k$ and randomly selects k examples from them. (4) Diverse nearest with BM25 filter (*DNN,BM*), filters out DNN examples not reaching the BM25 threshold. (5) *Random*, uniformly selects k random examples. (6) Mixed nearest neighbors (*MixedNN*), mixes the using of the NN and random retrieval strategies with the ratio of NN set to a.

Training with diverse retrieval strategies. Fig. 3 visualize the results of training with various retrieval strategies. We conduct inference with and without examples for each strategy, and set the retrieval strategy of inference the same as of training. The most straight forward method *NN* shows best performances, suggesting the benefits of semantically similar examples. *Random* strategy, though inferior to *NN*, also shows improvements, indicating that random examples might introduce some general information of NER taks to the model. Meanwhile, inference with examples does not guarantee improvements and often hurt performances. This may due to the differences of the annotation schema between the automatically constructed data and the human-annotated benchmarks.

Inference with out-domain examples. During inference, since examples from the automatically constructed data is not aligned with the domains and schemas of the human-annotated benchmarks, we refer to them as *out-domain examples*. Fig. 4 shows the results of inference with out-domain examples using diverse retrieval strategies. We use the model trained with NN strategy here. After applying example filtering such as BM25 scoring, inference with out-domain examples shows improvements compared to the baseline, suggesting the

need of example filtering when implementing RAG with out-domain examples.

Inference with in-domain examples. We explore the setting where a few *in-domain examples* are available for inference. We randomly sample an example pool of size N from the original training sets of the benchmarks, then retrieve k NN from this pool as in-domain examples. We also evaluate on *full* pool where the entire training set is used for retrieval. Results are shown in Fig. 5. In-domain examples show substantial improvements in Chinese. Meanwhile, sufficient in-domain examples are required for improvements in English. This indicates the benefits of providing sufficient in-domain examples for RAG.

Based on the above analysis, we suggest implementing on-demand RAG for inference after RA-IT. When sufficient in-domain examples are available, conduct RAG with similar examples to boost inference. When only out-domain examples are available, apply an example filtering method such as BM25 scoring for RAG, or simply conduct inference without examples.

4 Related Work

4.1 IE with LLMs

The main techniques studied in the area of IE with LLMs fall under advanced prompt designing(Guo et al., 2023; Xie et al., 2023; Wang et al., 2023a), instruction tuning (IT) (Sainz et al., 2024; Zhou et al., 2024; Li et al., 2024) and data augmentation (Josifoski et al., 2023; Zhang et al., 2023; Ma et al., 2023). Many of the prompt designing methods apply RAG to an off-the-shelf LLM to assist inference (Guo et al., 2023; Wan et al., 2023; Xie et al., 2024), which retrieves similar examples to provide more useful information for the LLM. Works of IT incorporate the information for IE into the LLMs through task-specific fine-tuning (Sainz et al., 2024; Zhou et al., 2024). Different from previous works, we explore retrieval augmented IT (RA-IT) for IE, with a focus on the open NER task.

Following UniNER (Zhou et al., 2024), we conduct investigations under the targeted distillation setting, since UniNER successfully distills the strong capability of ChatGPT in open NER into a smaller model without any human-annotated data. Other works of IT for IE, Sainz et al. (2024); Li et al. (2024) adopt the code-style instruction to fine-tune LLMs in effectively generating IE outputs through code generation. They are **orthogonal** to this work since the strategy of RA-IT can be integrated into various styles of instructions. Moreover, Zaratiana et al. (2023) integrated the strong capability of ChatGPT in open NER into smaller-scale bidirectional LMs (BiLMs) such as BERT (Devlin et al., 2019). How to integrate retrieval augmentation into the BiLMs frameworks is also worth exploring in future work.

4.2 Retrieval aware Fine-Tuning

Retrieval augmented generation (RAG) has achieved large improvements in diverse tasks with the off-the-shelf LLMs (Ram et al., 2023). Recent works has explored retrieval aware IT for LLMs (Jiang et al., 2023; Zhang et al., 2024). Jiang et al. (2023) pre-trains a retriever and LM jointly, then conducts few-shot fine-tuning on downstream tasks. Luo et al. (2023) instruction-tunes LMs with retrieved passages prepended to inputs. Zhang et al. (2024) retrieves both gold and distractor documents for IT to make the model resistant to unhelpful documents. Liu et al. (2024) explores context-enhanced IT to enhance model's capability for conversational QA over a given context. However, retrieval augmented and context-enhanced IT has remained unexplored in IE. We fill this gap and explore (RA-IT) on the task of open domain NER.

5 Conclusion

This paper explores RA-IT for open NER. We retrieve semantically similar examples to form the context-enhanced instruction data. RA-IT achieves consistent improvements across various data sizes in English and Chinese, suggesting the need of context-enhanced training. Thorough analysis verifies the benefits of semantically similar examples for training and the need of example filtering and in-domain examples for inference.

Limitations

This work faces the following limitations:

(1) Although the RA-IT strategy improves the open NER performance, it does not guarantee improvements when using RAG during inference. Applying some example filtering strategies and introducing in-domain examples alleviate this problem, but the effectiveness is till marginal. More advanced approaches of improving RA-IT models in conducting RAG for open NER are worth exploring.

(2) The investigation of data efficiency in this work is merely a small preliminary empirical study. However, data efficiency, such as selecting most influential and beneficial data is important for real-world applications of IE since it might effectively save computation and annotation costs.

Acknowledgements

This research is supported by Zhejiang Provincial Natural Science Foundation of China (LDT23F02023F02) and the National Natural Science Foundation of China (72350710798).

References

Moonshot AI. 2024a. Moonshot ai.

Zhipu AI. 2024b. Glm-4 releasing.

Anthropic. 2024. Claude 3 haiku: our fastest model yet.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

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

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *Preprint*, arXiv:2311.15653.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Hao Yang, and Tong Xiao. 2024. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191*.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

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

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

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, Long Bai, Wei Li, Yidan Liu, Pan Yang, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2024. Knowcoder: Coding structured knowledge into llms for universal information extraction. *Preprint*, arXiv:2403.07969.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. *Preprint*, arXiv:2401.10225.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *Preprint*, arXiv:2305.15225.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2023. STAR: Improving low-resource information extraction by structure-to-text data generation with large language models. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Hello gpt-4o.

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*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Qwen Team. 2024. Introducing qwen1.5.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

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. *Preprint*, arXiv: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. *Preprint*, arXiv:2304.08085.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lv, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for instruction tuning.

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

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. *Preprint*, arXiv:2311.08921.

Liang Xu, Yu tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, Caiquan Liu, and Xuanwei Zhang. 2020. Cluener2020: Fine-grained

named entity recognition dataset and benchmark for chinese. *Preprint*, arXiv:2001.04351.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer. *Preprint*, arXiv:2311.08526.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

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

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *Preprint*, arXiv:2403.10131.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

A Chinese Data Construction

Data construction prompt. Fig. 6 shows the prompt used for Chinese distillation data construction. We follow Zhou et al. (2024) to design the prompt for Chinese data construction. We adopt the data construction prompt of Pile-NER-type [3], since it shows the best performance as in (Zhou et al., 2024).

Figure 6: Data construction prompt for Chinese open domain NER.

Data processing. Following (Zhou et al., 2024), we chunk the passages sampled from the Sky corpus[4] to texts of a max length of 256 tokens and randomly sample 50K passages. Due to limited computation resources, we sample the first twenty files in Sky corpus for data construction, since the size of the entire Sky corpus is beyond the processing capability of our machines. We conduct the same data processing procedures including output filtering and negative sampling as in UniNER. Specifically, the negative sampling strategy for entity types, is applied with a probability proportional to the frequency of entity types in the entire constructed dataset.

Instruction data construction. The instruction tuning data for Chinese scenario is as shown in Table. 15.

B Diverse Teachers for Data Construction

We explored the effect of using diverse teachers for data construction. We also tried ensemble distillation: multiple teachers are used to annotate entities simultaneously, and the final annotation is acquired by majority voting on the answers from all teachers. The results are shown in Table 4

[3] https://huggingface.co/datasets/Universal-NER/Pile-NER-type

[4] https://huggingface.co/datasets/Skywork/SkyPile-150B

Model		Ontonotes 4	MSRA	Weibo	Boson	ClueNER	CMeEE	Ren.	Yidu	Avg.
ChatGPT	Teacher	29.7	41.4	30.3	46.7	44.8	43.2	34.3	34.9	38.1
	Student	48.0	52.6	38.0	52.2	42.4	41.0	48.8	49.0	**46.5**
Claude	Teacher	35.0	46.4	29.9	35.7	46.3	43.1	34.1	31.1	37.7
	Student	50.1	52.6	36.9	47.2	44.0	42.6	49.7	46.6	**46.2**
Moonshot	Teacher	43.4	51.6	27.5	51.6	49.6	43.1	41.8	34.7	42.9
	Student	50.8	49.7	34.9	54.3	43.5	43.0	48.2	52.1	**47.0**
GLM-4	Teacher	36.7	49.4	28.3	38.3	49.1	45.9	34.4	35.5	39.7
	Student	47.9	45.2	35.0	49.1	42.8	44.1	45.2	48.1	**44.7**

Table 4: The method of LLMs' extraction and LLM-guided SFT on Chinese dataset. "Teacher" refers to the results directly returned by instruct-LLM. "Student" refers to the results after SFT using the UniNER method by LLMs' results.

Language	Dataset	Labels	Train	Valid	Test
English	CrossNER_AI	13	100	350	431
	CrossNER_literature	11	100	400	416
	CrossNER_music	12	100	380	465
	CrossNER_politics	8	199	540	650
	CrossNER_science	16	200	450	543
	MIT Moive Review	12	9774	2442	2442
	MIT Restaurant Review	8	7659	1520	1520
Chinese	Ontonotes4	4	15724	4301	4346
	MSRA	3	46364	-	4365
	Weibo	4	1350	270	270
	Boson	6	1637	184	179
	ClueNER	10	10748	1343	-
	CMeEE	9	15000	5000	-
	Ren.	4	228616	28768	28885
	Yidu	6	1000	-	379

Table 5: Dataset statistic.

We investigated four representative powerful teacher models, ChatGPT (gpt-3.5-turbo-0125) (OpenAI, 2022), Claude 3 (claude-3-haiku) (Anthropic, 2024), Moonshot (moonshot-v1-8k) (AI, 2024a) and GLM-4 (glm-4) (AI, 2024b) for distilling open NER capability.

C More Details of Experimental Settings

C.1 Preliminary Study on Data Efficiency

We explore the impact of various data sizes: [0.5K, 1K, 5K, 10K, 20K, 30K, 40K, 50K]. For each data size, we randomly sample two sets of data and report the average.

C.2 Training

we use the training data Pile-NER released by Zhou et al. (2024), and we adopt the Pile-NER-type version[5], which shows better performance than Pile-NER-definition[6]. In our practice, we filter out samples with unparseable entity outputs in Pile-NER-type, which finally leaves 45K samples for actual experiments.

Through our experiments, we train models for 3 epochs with a batch size of 8 and a learning-rate of 5e-5. A cosine scheduler is adopted. Each experiment is run on one single A100 GPU.

C.3 Retrieval

For diverse nearest strategy, we set $K = 128$. For example filtering with BM25 scoring, we set the BM25 score threshold as 20.

[5]https://huggingface.co/datasets/Universal-NER/Pile-NER-type

[6]https://huggingface.co/datasets/Universal-NER/Pile-NER-definition

C.4 Benchmarks

For English, we adopt benchmarks CrossNER (Liu et al., 2021) and MIT-movie/restaurant (Liu et al., 2013). For Chinese, we collect eight benchmarks across diverse domains, include Ontonotes 4 (Weischedel et al., 2011), MSRA (Levow, 2006), Weibo (Peng and Dredze, 2015), ClueNER (Xu et al., 2020), CMeEE (Zhang et al., 2022), Yidu-S4k [7], Boson and PeopleDaily2014(abbreviated as 'Ren.' in the text and tables) [8] The following are our **sampling strategies** on evaluation data: For evaluated benchmarks, we sample 2000 examples for each test set for evaluation and keep the original test set with fewer than or slightly more than 2000 examples. For those datasets with only the training set and validation set publicly accessible, we randomly select half of the validation data as the test data and the other half as the new validation data.

D Full results on data size study

We provide the full results of preliminary study on data sizes. English dataset results are shown in Table 6. When the number of training data is less than 10K, the model performance improves significantly with data increasing. However, the results do not improve the performance after the number of data exceeds 20K. When applied to the Chinese datasets, the threshold increases to 30K in Table 7

E Extended Analysis

E.1 Varying Numbers of Examples

We keep the number of examples as 2 through our main experiments. Here we explore the impact of increasing the number of examples. We found that increasing the number of examples does not guarantee improvements. This is presumably because that the entire inputs get lengthy when the number of examples increases. And the very long input sequence is challenging for the 7B model to understand. The results for the Chinese datasets and the English datasets are shown in Table 11 and Table 8, respectively.

E.2 Ablation for Mixed NN

We conduct ablation experiments by mixing different sampling methods, constructing the training set using NN strategy and random strategy in various proportions. The samples were mixed at four different ratios of 0.2, 0.4, 0.6, and 0.8, respectively. Additionally, we both used two different strategies which inference with and without using examples. The experimental results for the Chinese and English datasets are shown in Table 12 and Table 13, respectively. The NN strategy performed the best overall, while the random sampling strategy did not contribute to the results.

F Examples of Instruction Data

We show examples of instruction data for English and Chinese in Table 14 and Table 15, respectively. For each sample, we used different strategies to find the k number of different examples in the training set. These prompts were formatted as "Text: sample text.\n Entity: [{entity text 1: entity type 1}, {entity text 2: entity type 2}]\n " for the large model to learn from. After the model learned from these examples, we had the model read the test sample text. Then, we asked a question for each entity category. The model's answers were organized as the information extraction results for the test samples.

G Case Study

We conduct case study to explore the advantages and disadvantages of the proposed RA-IT method. Table 9 are two cases that demonstrate that RA-IT benefits the long-tail entity types. We exclude the top 30% of frequent types and regard the remaining types as long-tail types. The entities in bold are long-tail types that are misclassified by vanilla-IT and corrected by RA-IT. These two cases are also commonsense-related.

Table 10 are some bad cases where RA-IT fails to improve. The entities in bold are vanilla-IT wrongly recognized, and RA-IT failed to improve. These professional entities in biomedical or AI domains require domain knowledge to be recognized. This shows that RA-IT benefits commonsense-related cases more than knowledge-seeking cases.

[7] http://old.openkg.cn/dataset/yidu-s4k

[8] People Daily 2014 and Boson datasets are available at https://github.com/hspuppy/hugbert/tree/master/ner_dataset.

	Data Size	Movie	Restaurant	AI	Literature	Music	Politics	Science	Avg.
ChatGPT	-	68.50	67.00	5.30	32.80	52.40	39.80	66.60	47.50
RA-IT	0.5K	50.92	41.50	47.61	54.27	54.15	54.78	52.16	50.77
	1K	44.88	39.74	50.21	56.14	55.97	56.28	54.01	51.03
	5K	44.87	42.72	52.87	59.00	60.47	59.35	58.36	53.95
	10K	49.81	41.47	53.78	60.99	63.79	60.84	61.47	56.02
	20K	50.14	42.17	57.07	62.02	65.43	61.92	63.35	**57.44**
	30K	47.21	40.84	58.15	63.11	65.33	62.64	64.46	57.39
	40K	45.89	40.34	56.34	62.48	64.71	62.12	63.58	56.49

Table 6: Impact of different dataset sizes on model performance in English scenario. Data size indicates the number of sampled data for prompt fine-tuning.

	Data Size	Ontonotes 4	MSRA	Weibo	Boson	ClueNER	CMeEE	Ren.	Yidu	Avg.
ChatGPT	-	29.70	41.36	30.25	46.65	44.75	43.16	34.25	34.90	38.13
RA-IT	0.5K	47.05	47.20	37.25	44.40	43.08	37.59	40.07	42.05	42.34
	1K	43.55	46.48	42.41	47.58	42.87	38.28	41.39	45.23	43.47
	5K	48.88	51.47	38.95	52.47	43.54	41.50	47.51	47.23	46.44
	10K	46.28	52.56	39.26	52.92	45.42	42.59	47.99	47.95	46.87
	20K	44.05	51.94	35.52	55.67	42.97	42.46	48.68	48.65	46.24
	30K	42.75	48.46	35.44	53.49	43.04	42.75	48.43	35.44	**48.71**
	40K	44.19	49.22	33.83	52.84	42.93	42.59	47.90	33.83	48.06

Table 7: Impact of different dataset sizes on model performance in Chinese scenario. Data size indicates the number of sampled data for prompt fine-tuning.

	#Example	Ontonotes 4	MSRA	Weibo	Boson	ClueNER	CMeEE	Ren.	Yidu	Avg.
ChatGPT	-	29.70	41.36	30.25	46.65	44.75	43.16	34.25	34.90	38.13
Vanilla IT	-	48.88	51.47	38.95	52.47	43.54	41.50	47.51	47.23	46.44
RA-IT	2	49.23	53.08	37.43	52.64	43.27	43.87	48.31	48.47	**47.04**
	4	48.80	53.32	35.50	50.65	43.41	43.82	48.34	48.86	46.59
	6	47.27	53.42	38.17	53.77	44.09	42.90	47.67	48.51	46.98
	8	52.20	50.56	40.18	50.10	43.69	39.17	44.53	49.45	46.24
	10	42.56	35.15	33.71	43.48	37.28	30.71	37.23	49.88	38.75

Table 8: Impact of different number of example on RA-IT performance of Chinese scenario. #Example column indicates the number of examples used for RA-IT

Case 1	
Input text	What is the theme song to stand by me?
Ground Truth	[**{'stand by me': 'title'}**]
Answer from Vanilla-IT	[**{'stand by me': 'song'}**, {'theme song': 'song'}]
Answer from RA-IT	[**{'stand by me': 'title'}**, {'theme song': 'song'}]
Case 2	
Input text	How many times has matt damon been jason bourne?
Ground Truth	[**{jason bourne: character}**, {matt damon: actor}]
Answer from Vanilla-IT	[{paul greengrass: director},**{jason bourne: title}**, {matt damon: actor}]
Answer from RA-IT	[**{jason bourne: character}**, {matt damon: actor}]

Table 9: Cases that RA-IT benefits the long-tail entity types. We exclude the top 30% of frequent types and regard the remaining types as long-tail types. The entities in bold are long-tail types that are misclassified by vanilla-IT and corrected by RA-IT. These two cases are also commonsense-related.

Case 1	
Input text	Viral TK phosphorylates aciclovir into its monophosphate form , which is subsequently phosphorylated to active aciclovir triphoshate by cellular kinase s, thus selectively inhibiting viral DNA polymerase.
Ground Truth	[**{Viral TK: enzyme}, {cellular kinase: enzyme}**, {DNA polymerase: enzyme}, {aciclovir triphoshate: chemical compound}]
Answer from Vanilla-IT	[**{Viral TK: scientist}, {cellular kinase s: scientist}, ...**]
Answer from RA-IT	[**{Viral TK: scientist}, {cellular kinase s: scientist}, ...**]
Case 2	
Input text	NIST also differs from Bilingual evaluation understudy in its calculation of the brevity penalty insofar as small variations in translation length do not impact the overall score as much.
Ground Truth	[**{NIST: metrics}, {bilingual evaluation understudy: metrics}**]
Answer from Vanilla-IT	[**{NIST: organization}, {bilingual evaluation understudy: organization}, ...**]
Answer from RA-IT	[**{NIST: organization}, {bilingual evaluation understudy: organization}, ...**]

Table 10: Cases where RA-IT fails to improve. The entities in bold are vanilla-IT wrongly recognized, and RA-IT failed to improve. These professional entities in biomedical or AI domains require domain knowledge to be recognized.

Method	#Example	Movie	Restaurant	AI	Literature	Music	Politics	Science	Avg.
ChatGPT	-	68.50	67.00	5.30	32.80	52.40	39.80	66.60	47.50
Vanilla-IT	-	52.87	59.00	60.47	59.35	58.36	44.87	42.72	53.95
RA-IT	2	52.61	60.01	63.04	60.02	58.91	50.26	45.75	**55.80**
	4	51.08	59.30	62.40	59.18	58.15	51.15	45.88	55.31
	6	48.79	54.46	55.75	54.62	50.93	54.79	46.81	52.31
	8	34.61	36.17	34.12	34.30	35.27	45.29	36.50	36.61
	10	22.89	30.35	22.03	30.24	26.41	38.31	35.80	29.43

Table 11: Impact of different number of example on RA-IT performance of English scenario. #Example column indicates the number of examples used for RA-IT

Method	ratio	#Exam.	Movie	Restaurant	AI	Literature	Music	Politics	Science	Avg.
ChatGPT	-	-	68.50	67.00	5.30	32.80	52.40	39.80	66.60	47.49
Vanilla-IT	-		44.87	42.72	52.87	59.00	60.47	59.35	58.36	53.95
RA-IT	0.2	0	45.65	45.25	51.89	58.36	62.00	59.26	58.21	54.37
		2	41.58	39.04	50.21	57.21	59.23	58.89	58.30	52.31
	0.4	0	46.81	45.62	51.72	58.92	62.42	59.39	58.62	54.78
		2	42.84	39.24	50.05	57.48	59.85	59.16	58.37	52.43
	0.6	0	48.56	45.05	51.5	58.89	61.47	59.05	57.18	54.53
		2	42.92	37.70	49.99	57.41	59.01	58.76	57.84	51.95
	0.8	0	47.82	45.49	52.29	58.81	61.90	59.61	58.70	54.95
		2	41.60	37.72	49.89	57.63	59.62	58.89	57.69	51.86
	1	0	53.20	47.36	52.50	60.86	63.13	60.77	61.02	**56.98**
		2	43.24	38.56	49.71	58.29	60.74	59.94	59.10	52.80

Table 12: Results of mixing random strategy with NN strategy at different ratios in English dataset. 'ratio' indicates the proportion of NN strategy in the total number of samples while training. When ratio=1, all samples are from NN strategy. '#exam.' indicates whether example data was added to prompts during testing, with 0 indicating no addition, and 2 indicating 2 examples retrieved by NN strategy added to the test examples.

Method	ratio	#Exam.	Ontonotes 4	MSRA	Weibo	Boson	ClueNER	CMeEE	Ren.	Yidu	Avg.
ChatGPT	-	-	29.70	41.36	30.25	46.65	44.75	43.16	34.25	34.90	38.13
Vanilla-IT	-	-	48.88	51.47	38.95	52.47	43.54	41.50	47.51	47.23	46.44
RA-IT	0.2	0	49.36	52.53	37.57	52.42	42.86	43.44	48.45	48.03	46.83
		2	47.93	49.97	37.30	51.40	43.10	41.71	47.97	47.35	45.84
	0.4	0	49.03	52.71	37.81	52.52	42.83	43.53	48.33	48.04	46.85
		2	48.03	50.30	37.33	51.43	43.17	41.59	47.62	46.99	45.81
	0.6	0	48.91	52.43	37.01	51.35	42.59	43.68	48.24	48.04	46.53
		2	47.83	49.85	37.62	53.03	42.92	41.77	47.79	47.87	46.09
	0.8	0	49.25	52.86	37.27	52.88	43.15	43.92	48.32	48.35	47.00
		2	48.04	49.95	38.12	50.61	43.04	41.58	47.81	47.63	45.85
	1	0	50.46	53.73	38.10	54.36	43.70	43.78	48.65	48.87	**47.71**
		2	47.96	50.90	40.00	53.68	43.98	42.09	47.75	48.56	46.87

Table 13: Results of mixing random strategy with NN strategy at different ratios in Chinese dataset. 'ratio' indicates the proportion of NN strategy in the total number of samples while training. When ratio=1, all samples are from NN strategy. '#exam.' indicates whether example data was added to prompts during testing, with 0 indicating no addition, and 2 indicating 2 examples retrieved by NN strategy added to the test examples.

Role	Conversation
Human	Here are some examples of named entity recognition: Text: 50 Top B2B Marketing Influencers 2017. It's October and you know what that means? Its B2B Marketing influencer speaker list time again. One of my all-time favorite conferences is MarketingProfs B2B Forum in Boston and for the past few years. I've had some fun listing out a top list of speakers ranked by influence around the topic of "B2B marketing". As usual, I used the influencer marketing platform Traackr to import the list of speakers from #mpb2b 2017 and rank them according to a combination of topical resonance and relevance as well as network reach related to "b2b marketing". Of course, use of their platform in this way is like 1% of what Traackr can do. I imagine they cringe every time I use their robust tool for such a simple list - but hey, they provide me with access and I use the tool as I see fit. To clarify, my agency TopRank Marketing is also a paying customer of the Traackr platform for clients, where it is used in support of B2B influencer marketing programs for brands like SAP, BMC Software, McKesson and others in ways that are more in line with the platform's capabilities. This is a legit list that recognizes people creating content around B2B marketing that resonates with their social following. Entity: [{'TopRank Marketing': 'agency'}, ... {'mpb2b 2017': 'event'}] Text: How to know and choose online games: differences between current and potential players. This study investigated how different adolescent players acquire game information and the criteria they use in choosing online games and found that (1) current players generally use comprehensive information sources more than potential players do; (2) current players rely on free trials and smooth display of motion graphics as choice criteria more than potential players do; (3) potential players rely on the look of advertisements more than current players do; (4) both current and potential players most likely use word-of-mouth and gaming programs on TV as information sources; and (5) endorser attractiveness is ranked the least important among six choice criteria by both current and potential players. Entity: [{'online games': 'Product'}, ... {'potential players': 'Person'}]
LLM	I've read these examples.
Human	Text: For brands looking to get into the rising world of esports, sponsoring live streamers on twitch is a popular choice. However, it appears competition may be heating up for endorsements from the top 0.2% of talent. These superstars typically represent the top 50 channels online at any given time and typically have lined up between 2 and 8 simultaneous sponsors leading to a crowded space for brands looking to get involved with esports. Meanwhile, brands able to engage popular streamers below the superstar level, are on average the channel's only sponsor. At Endorse.gg we provide an integrated analytics and engagement platform to help manage large scale and largely exclusive campaigns across a much greater number of these smaller but high quality broadcasters. Source: Analysis of live streamers on twitch.tv over the course of a week. "Top 10 channels" refers to the top 10 online at any given moment. Based on an analysis of currently live channels at several points over the course of a week, we see that the top 10 channels by number of viewers typically command $\tilde{1}$5% of the total twitch viewership and have $\tilde{3}$ actively endorsed brands. Meanwhile, channels #11-50 capture an additional 25% of the views while promoting 2.2 brands on average.
LLM	I've read this text.
Human	What describes "organization" in the text?
LLM	["brands", "twitch", "Endorse.gg"]

Table 14: A English sample of prompts and LLM's responses. In the Role column, "Human" indicates the prompt and "LLM" indicates the LLM response.

Role	Conversation
Human	以下是一些命名实体识别的例子： 文本：创刊以来，办刊质量和水平不断提高，主要栏目设置有：论著、社区卫生管理、综述讲座、社区疾病预防、社区健康教育、卫生政策、社区计划生育等。 期刊收录：知网收录(中)上海图书馆馆藏维普收录(中)国家图书馆馆藏万方收录(中) 《四川医学》杂志在全国影响力巨大，创刊于1980年，公开发行的月刊杂志。创刊以来，办刊质量和水平不断提高，主要栏目设置有：基金论文、实验研究、临床研究、医学进展、自学辅导站、卫生管理、检验医学、调查报告等。 期刊收录：知网收录(中)万方收录(中)国家图书馆馆藏维普收录(中)CA 化学文摘(美)上海图书馆馆藏统计源核心期刊(中国科技论文核心期刊) 《中国医学工程》杂志在全国影响力巨大，创刊于2002年，公开发行的月刊杂志。 实体：['期刊收录': '事件', ... ,'检验医学': '类别'] 文本：《中国骨伤》杂志是我国大陆地区第一个被美国《医学索引》（IndexMedic 本刊由国家中医药管理局主管，中国中医科学院中药研究所和中华中医药学会主办，是以报道、介绍中医药实验方剂研究为主旨的专业性学术期刊，创刊于1995年10月。本刊办刊方针以提高为主、提高与普及相结合。设置栏目有中医临床、药理、毒理、药物代谢、药剂与炮制、化学与分析、资源与鉴定、综述、专论、数据挖掘、中医传承、标准与规范等，交流方剂及其主要组成药物的研究成果与最新进展。读者对象是从事中、 《中华中医药杂志》原名《中国医药学报》，是中国科学技术协会主管、中华中医药学会主办的国家级高级中医药学术期刊，是中国科学技术协会所属的自然科技期刊中反映中医药研究进展的中医学科杂志。 实体：['国家级高级中医药学术期刊': '期刊', ... ,'毒理': '栏目']
LLM	我已读完这些例子。
Human	文本：《福建医药》是综合性医药学术期刊，综合因子为：0.164，期刊级别为：省级期刊。福建医药及时报道国内外医药卫生科研成果及防治疾病经验，反映其进展及水平，为省内外各级医药卫生技术人员提供学术交流平台。 期刊收录：国家图书馆馆藏上海图书馆馆藏维普收录(中)知网收录(中)万方收录(中)CA 化学文摘(美) 《机械工程与自动化》（CN：14-1319/TH）是一本有较高学术价值的大型双月刊，自创刊以来，选题新奇而不失报道广度，服务大众而不失理论高度。颇受业界和广大读者的关注和好评。 《齐齐哈尔大学学报》是省级期刊、荣获首届全国优秀社科学报，综合影响因子为0.074。
LLM	我已读完这段文本。
Human	文本中属于"组织"的实体有哪些?
LLM	[]
Human	文本中属于"期刊"的实体有哪些?
LLM	["齐齐哈尔大学学报", "机械工程与自动化", "福建医药"]
Human	文本中属于"收录机构"的实体有哪些?
LLM	["上海图书馆", "维普", "万方", "CA 化学文摘", "知网", "国家图书馆"]

Table 15: A Chinese sample of prompts and LLM's responses. In the Role column, "Human" indicates the prompt and "LLM" indicates the LLM response. The format of the Chinese prompts is the same as the English.