Improving Open Information Extraction with Large Language Models: A Study on Demonstration Uncertainty

Anonymous authors
Paper under double-blind review

Abstract

Open Information Extraction (OIE) task aims at extracting structured facts from unstructured text, typically in the form of (subject, relation, object) triples. Despite the potential of large language models (LLMs) like ChatGPT as a general task solver, they lag behind state-of-the-art (supervised) methods in OIE tasks due to two key issues. First, LLMs struggle to distinguish irrelevant context from relevant relations and generate structured output due to the restrictions on fine-tuning the model. Second, LLMs generate responses based on probability, which makes the predicted relations lack confidence. In this paper, we assess the capabilities of LLMs in improving the OIE task. Particularly, we propose various in-context learning strategies to enhance LLM's instruction-following ability and a demonstration uncertainty quantification module to enhance the confidence of the generated relations. Our experiments on three OIE benchmark datasets show that our approach holds its own against established supervised methods, both quantitatively and qualitatively.

1 Introduction

Open Information Extraction (OIE) involves the identification and extraction of novel relations and their components (e.g., subjective, action, objective, and adverbs) from unstructured text. It enables the creation of large-scale knowledge graphs from diverse sources (Wang et al., 2018), aiding in tasks like question answering (Ling et al., 2023a), knowledge-augmented reasoning (Chowdhury et al., 2023), and semantic search (Niklaus et al., 2018). As a frontier technology, ChatGPT (Ouyang et al., 2022) and other large language models (LLMs) (Ling et al., 2023b; Chang et al., 2024; Bai et al., 2024) excel at comprehending and producing a wide variety of intricate natural language constructs. Therefore, they naturally present a promising solution for solving the OIE task without requiring substantial training.

The conventional OIE methods are trained on labeled data, where each entity and its relations are explicitly annotated. This allows them to learn precise patterns and directly map input to specific output tags, resulting in high accuracy. Despite the potential of LLMs like ChatGPT as a general task solver, they lag behind tagging-based methods in OIE tasks due to two key issues (Ling et al., 2023b). First, LLMs as a generative model are trained to generate human-like text and not specifically for information extraction. While they have a broad understanding of language and can generate coherent responses, they may not be as accurate or consistent in extracting specific pieces of information from the text as supervised models trained specifically for that task. Second, the responses generated by LLMs are based on the input prompt and are probabilistic, which can result in outputs with lower confidence. This lack of confidence can engender inconsistencies, such as the same relation being extracted differently in varying contexts or not being extracted at all in certain instances. Furthermore, this diminished confidence can lead to the extraction of incorrect or irrelevant relations, thereby reducing confidence in interpreting the extracted relations.

While zero-shot LLMs cannot solve complex OIE problems solely with original task instructions, there are a few attempts (Lu et al., 2023; Han et al., 2023; Wei et al., 2023) trying to solve OIE with LLMs in different ways. A recent method (Wei et al., 2023) is proposed to tackle the information extraction with ChatGPT interactively by decomposing the framework into several parts and then combining the results of each round into a final structured result, but they can only handle OIE tasks with fixed relations. Lu et al. also proposed to leverage instruction tuning to enhance LLMs
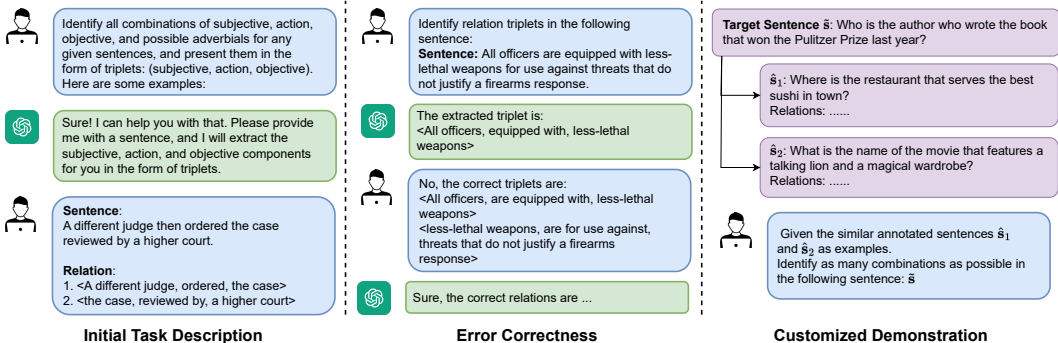
Figure 1: The framework of the proposed method consists of 1) providing an initial task description; 2) setting up a quiz to enhance ChatGPT's understanding of the OIE task, and 3) customized demonstration selection.

on a hand-crafted dataset, however, their method has to be extensively fine-tuned on hand-crafted datasets (Lu et al., 2023). Another recent approach (Han et al., 2023) focuses on investigating the capability of ChatGPT in the OIE task from various aspects. However, none of the existing works have considered enhancing the robustness and confidence of the response. In this work, we summarize our key contributions as follows: (1) We propose a novel framework that allows LLMs to solve OIE tasks with various few-shot demonstration strategies without extensive fine-tuning; (2) We include an uncertainty quantification module to increase the confidence of the predicted answers. (3) A series of experiments have shown the effectiveness of the proposed method as well as each component in the proposed framework.

2 APPROACH

In this work, we focus on Relational Triplet Extraction, designed to identify entities and their relationships from a given sentence. To utilize LLMs in the task, we incorporate task-specific instructions, few-shot demos, an error correction mechanism, and an uncertainty quantification mechanism. This approach facilitates the generation of coherent and structured responses for the OIE task.

2.1 RELATION EXTRACTION WITH IN-CONTEXT LEARNING

Problem Formulation. Formally, given a sentence as a sequence of tokens/words $\tilde{\mathbf{s}} = <w_1, w_2, \cdots, w_n >$, Relational Triplet Extraction requires to output a list of tuples $T = [T_1, T_2, \cdots]$ with the i-th tuple $T_i = <w_s, p_i, w_o >$ representing a fact in the source sequence, where p denotes the predicate in T_i, w_s and w_o are the subjective and objective entities of T_i, respectively. As an autogressive generation model, ChatGPT outputs T based on the input prompt (i.e., task instruction), and the response is directly determined by designing a suitable prompt/instruction.

Initial Instruction Crafting. Task-specific instructions are essential for Open Information Extraction, particularly for the Relational Triplet Extraction task, as they guide the model to navigate the complexity and ambiguity of unstructured natural language and output structured responses for the task. Specifically, we incorporate a unified workflow with chain of instructions to guide the model step-by-step. As demonstrated in Figure 1, ChatGPT is kick-started by prompted with an *Initial Task Description* and a few demonstrations.

Error Correction. The error correction mechanism refines ChatGPT's understanding of the OIE task and enhances response accuracy. After introducing easy examples to ChatGPT, a quiz, acting as a validation tool, is set. As depicted in Figure 1, ChatGPT is tasked with extracting relational triplets from several sentences without prior correct annotations. Subsequent provision of correct answers helps rectify potential errors or incomplete responses. Empirically, this step bolsters ChatGPT's performance on actual OIE tasks, improving accuracy and dependability.

Few-shot Demonstration Selection. To better enable ChatGPT to generate responses consistent with the task description, we incorporate few-shot examples sampled from the training set that are similar to the target sentence $\tilde{\mathbf{s}}$. Specifically, given a training set $\mathcal{S}$ consisting of annotated sentences, we aim to retrieve a small subset $\hat{\mathcal{S}} \subseteq \mathcal{S}$ of structurally similar sentences with the target sentence $\tilde{\mathbf{s}}$. For example, if the target sentence $\tilde{\mathbf{s}}$ is an interrogative with an attribu-

	CaRB			OIE2016			ReOIE		
	PR	RE	F1	PR	RE	F1	PR	RE	F1
Rule-based: OpenIE4	-	-	48.0	-	-	60.0	-	-	68.3
Tagging-based: SpanOIE	60.9	41.6	49.4	-	-	69.4	79.7	74.5	**77.0**
Generation-based: IMOJIE	**64.7**	45.6	**56.8**	-	-	-	88.1	67.1	76.2
LLaMA-2-13B w/ Fixed Demo	41.6	28.0	33.4	37.5	30.6	33.7	31.2	18.1	22.9
LLaMA-2-13B w/ Selected Demo	43.2	30.4	35.7	40.7	32.6	36.2	32.3	21.3	25.7
LLaMA-2-13B w/ Selected Demo & Uncertainty	44.4	30.6	36.2	41.1	33.5	36.9	47.9	34.5	40.1
LLaMA-2-70B w/ Fixed Demo	43.7	51.0	47.1	57.5	**71.3**	63.7	60.9	70.4	65.3
LLaMA-2-70B w/ Selected Demo	55.4	47.1	50.9	68.4	61.6	64.8	62.1	72.5	66.9
LLaMA-2-70B w/ Selected Demo & Uncertainty	56.3	47.5	51.5	**68.9**	63.0	65.8	69.4	65.7	67.5
GPT-3.5-Turbo	41.0	37.5	39.1	41.5	51.3	45.9	35.8	20.2	25.9
GPT-3.5-Turbo w/ Fixed Demo	53.8	48.6	51.1	60.6	64.5	62.5	64.5	71.4	67.8
GPT-3.5-Turbo w/ Selected Demo	54.7	49.5	52.0	59.5	67.0	63.1	50.5	79.2	61.6
GPT-3.5-Turbo w/ Selected Demo & Uncertainty	54.3	50.1	52.1	61.6	69.1	65.1	65.8	70.2	67.9

Table 1: The performance of each approach on three popular benchmarks CaRB, OIE2016, and ReOIE with multiple partial matching strategies. The best results is highlighted with bold and the second best is highlighted with underline. The results missing in the literature are marked as -.

Sentence	Although in Flanders, the Flemish Region assigned all of its powers to the Flemish Community, the Walloon Region remains in principle distinct from and independent from the French Community, and vice-versa.
GPT-3.5-Turbo	1. (the Flemish Region, remains, in principle distinct from and independent from the French Community) 2. (Walloon Region, remains, in principle distinct from and independent from the French Community)
GPT-3.5-Turbo + Selected Demo	1. (the Flemish Region, assigned, all of its powers to the Flemish Community) 2. (the Walloon Region, remains, distinct from and independent from the French Community) 3. (the Walloon Region, remains, in principle distinct from and independent from the French Community) 4. (the French Community, is, distinct from and independent from the Walloon Region)
GPT-3.5-Turbo + Selected Demo + Uncertainty	1. (the Flemish Region, assigned, all of its powers to the Flemish Community) 2. (the Walloon Region, remains in principle distinct from, the French Community) 3. (the Walloon Region, remains independent from, the French Community) 4. (the French Community, is, distinct from and independent from the Walloon Region)
Golden Standard	1. (the Flemish Region, assigned, all of its powers) 2. (the Walloon Region, remains in principle distinct from, the French Community) 3. (the Walloon Region, remains independent from, the French Community)

Table 2: The extracted triplets with different demonstration strategies against the Golden Standard.

tive clause: `Who is the author who wrote the book that won the Pulitzer Prize last year?`, then the ideal structurally similar sentences from the training set should be: `Where is the restaurant that serves the best sushi in town?`. Each testing instance is given customized in-context learning demonstrations rather than fixed ones. Structural similarity is computed using cosine similarity between sentence latent embeddings, obtained through an Instruction-Fine-tuned Language Model (Su et al., 2022).

2.2 Demonstration Uncertainty

After completing the in-context training steps with few-shot demonstration selections and error correctness, the performance can already go beyond the original zero-shot ChatGPT. However, the autoregressive generation process's inherent probabilistic characteristics may impair confidence in the predictions. In this study, we introduce an innovative module designed to quantitatively assess the uncertainty associated with the predicted output, thereby bolstering confidence in the output.

In this work, we leverage an ensemble method to quantify the uncertainty in ChatGPT's responses. Specifically, after obtaining the structural similar annotated sentence set $\hat{\mathcal{S}} \subseteq \mathcal{S}$, we sample a list of small subset $[\hat{\mathcal{S}}_i], \hat{\mathcal{S}}_i \sim \hat{\mathcal{S}}$. We iteratively allow ChatGPT to generate answers of the target sentence $\tilde{s}$ by employing $\hat{\mathcal{S}}_i$ as an in-context learning example. Moreover, we leverage the prompt: "Identify **as many combinations as possible** in the following sentence: $\tilde{s}$" to ask ChatGPT to generate relational triplets even if it has low confidence in the triplet. We collect all the generated relational triplets from each sampled $\hat{\mathcal{S}}_i$ into a list $T_{\tilde{s}}$ and calculate the *Demonstration Uncertainty* $U_{\tilde{s}}$ of each $T_i \in T_{\tilde{s}}$ as: $U_{\tilde{s}} = \{u_i \mid 1 - \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{T_j = T_i}\}$, where N denotes the total number of elements in $T_{\tilde{s}}$ and $\mathbb{1}$ is an indicator function counting each element's occurrence. Intuitively, if the uncertainty score u_i is high, then it denotes the corresponding T_i appears less frequently in $T_{\tilde{s}}$, and vice versa. In this work, we adopt a threshold k to filter high uncertain T_i, i.e., $T_{\tilde{s}} = \{T_i \mid u_i \geq k\}$. Finally, based on the feedback from ChatGPT, we form the list of triplets $T_{\tilde{s}}$ to perform the evaluation.

3   EXPERIMENT

We utilize two families of LLMs, i.e., LLaMA-2 13B & 70B and GPT-3.5-TURBO. The framework is evaluated on three OIE benchmarks: (1) CaRB (Bhardwaj et al., 2019); (2) OIE2016 (Stanovsky & Dagan, 2016); and (3) ReOIE (Zhan & Zhao, 2020). OIE systems are typically evaluated by comparing the extractions with the gold set in each dataset and commonly used measures are Precision, Recall, and F1 scores. We follow the matching function proposed in each dataset.

Baselines. We adopt a list of recent OIE methods for comparison. (1) OpenIE4 (Mausam, 2016) is a rule-based relational extraction method. (2) SpanOIE (Zhan & Zhao, 2020) is a supervised method that directly predicts tokens' classes. (3) IMOJIE (Kolluru et al., 2020) generates the answer directly. Other than supervised methods, we incorporate different versions of our method using different LLMs. Specifically, (4) GPT-3.5-TURBO represents the vanilla version of the ChatGPT without any in-context learning and uncertainty quantification. (5) LLaMA-2 W/ FIXED DEMO and GPT-3.5-TURBO W/ FIXED DEMO denote a fixed three sentences with annotated relational triplets are provided as a demonstration. (6) LLaMA-2 W/ SELECTED DEMO and GPT-3.5-TURBO W/ SELECTED DEMO denotes the framework with customized in-context learning examples. (7) LLaMA-2 W/ SELECTED DEMO & UNCERTAINTY and GPT-3.5-TURBO W/ SELECTED DEMO & UNCERTAINTY denotes the final framework with considering all components. Note that we adopt two versions of LLaMA-2 model with different parameter size: 13B and 70B.

3.1   DISCUSSION

The experiment results are depicted in Table 1, and we draw a few conclusions from the result. First, the gap between the zero-shot LLM (GPT-3.5-TURBO) and the best methods across three datasets is around 30% in F1 score, which is reasonable since all SOTA methods are trained on corresponding datasets. Second, although few-shot LLM approachs are still not comparable to supervised fine-tuned methods in each dataset, the performance gap between the best LLM approach (i.e., GPT-3.5-TURBO W/ SELECTED DEMO & UNCERTAINTY) and the supervised methods are very little (on average 6% across three datasets). Third, with the growth of the parameter size, the few-shot OIE accuracy of LLMs also increases (LLaMA-2-13B< LLaMA-2-70B ≈ GPT-3.5-TURBO). Even though we don't know the exact parameter size of GPT-3.5-TURBO, the performance of the state-of-the-art open source LLM LLaMA-2-70B is nearly the same as GPT-3.5-TURBO. Furthermore, to mitigate the observed performance gap between the zero-shot LLMs and supervised methods, we propose several techniques to enhance the model's understanding of the task and increase the trustworthiness of the prediction along the way. Taking the GPT-3.5-TURBO as an example, by adding a few demonstration examples, GPT-3.5-TURBO W/ FIXED DEMO surpasses the zero-shot ChatGPT around 15% in F1 score across all datasets. By customizing the demonstrations and involving the uncertainty quantification module, the framework can achieve competitive and sometimes better results with state-of-the-art methods, the results are consistent in both 13B and 70B models.

3.2   CASE STUDY

We randomly pick a test instance in the OIE2016 dataset and let each ablated model generate a corresponding response. GPT-3.5-TURBO produced a correct number of relations, but it was incorrectly understood one concept stated in the original sentence. By adding selected demonstration examples, GPT-3.5-TURBO+SELECTED DEMO provided more accurate interpretations but still over-interpreted the independence of the French Community from the Walloon Region. Finally, considering the GPT-3.5-TURBO+SELECTED DEMO+UNCERTAINTY provided the most accurate interpretation compared to the golden standard.

4   CONCLUSION

We investigate the capability of LLMs being a zero/few-shot OIE system. We incorporate various in-context learning strategies to increase LLM's understanding of the task and allow it to generate structured output following the instruction. To further enhance the confidence of the generated output, we design an uncertainty quantification module to filter low-confident predictions. The proposed framework can achieve competitive results with methods that are extensively trained on each dataset. This work can potentially serve as a starting point to freely mine entities and relations for constructing domain-specific knowledge bases (Cui et al., 2023).

REFERENCES

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, et al. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*, 2024.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. CaRB: A crowdsourced benchmark for open IE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Jiayu Chang, Shiyu Wang, Chen Ling, Zhaohui Qin, and Liang Zhao. Gene-associated disease discovery powered by large language models. *arXiv preprint arXiv:2401.09490*, 2024.

Tanmoy Chowdhury, Chen Ling, Xuchao Zhang, Xujiang Zhao, Guangji Bai, Jian Pei, Haifeng Chen, and Liang Zhao. Knowledge-enhanced neural machine reasoning: A review. *arXiv preprint arXiv:2302.02093*, 2023.

Hejie Cui, Jiaying Lu, Shiyu Wang, Ran Xu, Wenjing Ma, Shaojun Yu, Yue Yu, Xuan Kan, Chen Ling, Joyce Ho, et al. A survey on knowledge graphs for healthcare: Resources, applications, and promises. *arXiv preprint arXiv:2306.04802*, 2023.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*, 2023.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Soumen Chakrabarti, et al. Imojie: Iterative memory-based joint open information extraction. *arXiv preprint arXiv:2005.08178*, 2020.

Chen Ling, Xuchao Zhang, Xujiang Zhao, Yifeng Wu, Yanchi Liu, Wei Cheng, Haifeng Chen, and Liang Zhao. Knowledge-enhanced prompt for open-domain commonsense reasoning. *AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making*, 2023a.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Tianjiao Zhao, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*, 2023b.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. Pivoine: Instruction tuning for open-world information extraction. *arXiv preprint arXiv:2305.14898*, 2023.

Mausam Mausam. Open information extraction systems and downstream applications. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, pp. 4074–4077, 2016.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*, 2018.

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

Gabriel Stanovsky and Ido Dagan. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2300–2305, 2016.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022.

Chengbin Wang, Xiaogang Ma, Jianguo Chen, and Jingwen Chen. Information extraction and knowledge graph construction from geoscience literature. *Computers & geosciences*, 112:112–120, 2018.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*, 2023.

Junlang Zhan and Hai Zhao. Span model for open information extraction on accurate corpus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9523–9530, 2020.