Enhancing Document-level Argument Extraction with Definition-augmented Heuristic-driven Prompting for LLMs

Tongyue Sun[1] and Jiayi Xiao[2,3]

[1]School of Engineering and Informatics, University of Sussex, Brighton, UK
[2]International Business School Suzhou, Xi’an Jiaotong-Liverpool University, Suzhou, China
[3]Management School, University of Liverpool, Liverpool, UK

Abstract

Event Argument Extraction (EAE) is pivotal for extracting structured information from unstructured text, yet it remains challenging due to the complexity of real-world document-level EAE. We propose a novel Definition-augmented Heuristic-driven Prompting (DHP) method to enhance the performance of Large Language Models (LLMs) in document-level EAE. Our method integrates argument extraction-related definitions and heuristic rules to guide the extraction process, reducing error propagation and improving task accuracy. We also employ the Chain-of-Thought (CoT) method to simulate human reasoning, breaking down complex problems into manageable sub-problems. Experiments have shown that our method achieves a certain improvement in performance over existing prompting methods and few-shot supervised learning on document-level EAE datasets. The DHP method enhances the generalization capability of LLMs and reduces reliance on large annotated datasets, offering a novel research perspective for document-level EAE.

1 Introduction

Event Argument Extraction (EAE) is a task within the domain of Natural Language Processing (NLP), focusing on the identification of relevant information pertaining to specific events from textual data. The majority of previous studies posit that events are articulated solely within a single sentence, hence their primary focus has been on sentence-level information extraction (Chen et al., 2015; Liu et al., 2018; Zhou et al., 2021). However, in real-life contexts, events are often narrated through complete documents composed of multiple sentences, such as news reports or medical records, an area that remains to be thoroughly investigated. Document-level EAE commonly relies on manual domain and pattern annotation for supervised learning models (Xiang and Wang, 2019; Lin et al., 2020; Li et al., 2022; Liu et al., 2022; Hsu et al., 2022; Liu et al., 2023). While this method is effective, it requires substantial labeling work. Considering the inherent complexity of document-level EAE, this is particularly burdensome and costly.

With the continuous evolution of Large Language Models (LLMs), their demonstrated potential has positioned them as formidable competitors to traditional methods in the field of EAE. For instance, InstructGPT (Ouyang et al., 2022) and ChatGLM (Du et al., 2022) have excelled in diverse downstream applications such as dialogue systems and text summarization generation through meticulously crafted instructions. Furthermore, recent studies (Lin et al., 2023; Zhou et al., 2024) have expanded the application of LLMs in complex tasks like event extraction by ingeniously constructing prompts, highlighting the broad prospects of LLMs in the EAE domain.

In prior research, pre-trained and fine-tuned models have exhibited deficiencies in generalization capabilities, largely constrained by the high costs of annotation and the risks of error propagation. The domain of document-level event argument extraction faces significant challenges, with the scarcity of high-quality datasets and the models’ insufficient generalization to unseen events being the primary bottlenecks. In contrast to the traditional reliance on vast corpora, the incorporation of In-Context Learning (ICL) within LLMs has emerged as a transformative approach (Brown et al., 2020; Zhou et al., 2022, 2023; Wang et al., 2024). ICL adeptly diminishes the necessity for extensive datasets by leveraging a modest collection of examples, serving as illustrative prompts for both inputs and outputs. This approach not only enhances the models’ adaptability but also significantly amplifies their proficiency in tackling tasks across a spectrum of novel and unseen instances. Heuristics are defined as *′a high-level rule or strategy for inferring answers to a specific task.′* and play a crucial role in human cognition. Humans use heuristics as an effective cognitive pathway, which often leads to more accurate reasoning than complex methods (Gigerenzer and Gaissmaier, 2011; Hogarth and Karelaia, 2007; Zhou et al., 2024). In ICL, heuristics are used to select or design examples (demonstrations) that can guide the model to make correct predictions(Zhou et al., 2024). By using examples generated by the model itself as context, the reliance on large-scale training datasets can be reduced, enhancing the model's adaptability. The performance of ICL is highly sensitive to specific settings, necessitating the selection of appropriate contextual information and the optimization of the model's training process. This includes the choice of prompt templates, the selection of context examples, and the order of examples (Zhao et al., 2021; Lu et al., 2022), as well as the selection of examples and the format of inference steps (Zhang et al., 2022b; Fu et al., 2022; Zhang et al., 2022a), which collectively impact the application of ICL on LLMs.

The Chain-of-Thought (CoT) (Wei et al., 2022) stands as an augmented prompting technique, widely recognized for its efficacy across a spectrum of tasks that demand sophisticated reasoning. CoT has proven particularly adept at tackling complex reasoning challenges, encompassing arithmetic and commonsense reasoning (Cobbe et al., 2021; Wei et al., 2022). However, its effectiveness is notably constrained in non-reasoning scenarios. When applied to tasks that do not inherently require reasoning, the CoT method risks simplifying the multi-step reasoning process into a potentially inadequate single-step, thereby undermining its full potential (Shum et al., 2023; Zhou et al., 2024). Consequently, there is a compelling need to devise specialized prompting strategies tailored for non-reasoning tasks. These strategies should be crafted to address the unique demands of such tasks, ensuring that the models maintain their robust performance across the diverse landscape of language processing challenges.

In this paper, we introduce a suite of innovative contributions aimed at advancing Event Argument Extraction and addressing the limitations of existing methods:

Definition-augmented Heuristic-driven Prompting Method. We improved the prompting heuristic method by incorporating argument extraction related definitions prompting and identified arguments. Utilizing inputs that include document content, task definitions, argument extraction rules, and identified event types and triggers, we constructed a definition-driven heuristic ICL. This method can process new situations (new classes) by analogy with known situations (known classes), effectively reducing error propagation and improving task accuracy. It provides a structurally complete and well-defined framework for events and arguments, incorporating necessary constraints. This not only improves the precision of extraction but also offers the model a richer and more consistent reference benchmark.

Chain-of-Thought Method. We employed the Chain-of-Thought method, guiding the model to incremental reasoning by providing coherent examples. These examples demonstrate how to break down complex problems into more manageable sub-problems and enhance the model's reasoning capabilities by simulating the human thought process.

Optimized Prompt Length. For the document-level Event Argument Extraction task, we fine-tuned the prompt length to enhance overall extraction performance. Such adjustments ensure that the token limit of LLMs is not exceeded. The prompt contains sufficient information while avoiding efficiency decline due to excessive length.

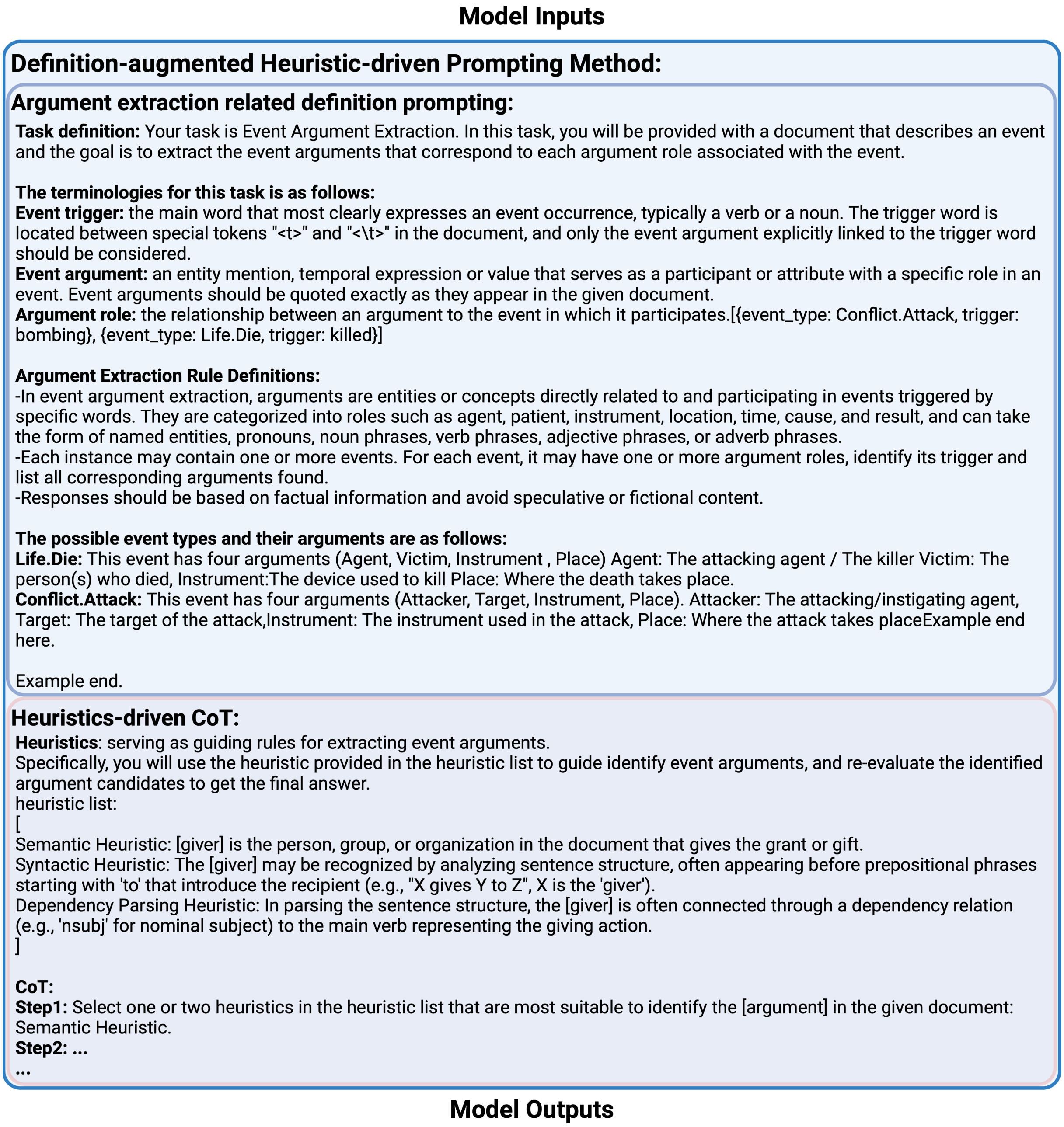
We propose new perspectives and methods, solving the example selection problem from the new perspective of Definition-Enhanced Prompting Heuristic Method, promoting explicit heuristic learning in ICL. The aim is to build more robust and adaptable prompting methods suitable for Event Argument Extraction. By implementing proof, it effectively improves task performance, enhances the model's ability to grasp the complex relationships between events and arguments, and contributes to further improving the capabilities of LLMs in EAE tasks.

2 Approach

We propose Definition-augmented Heuristic-driven Prompting Method for enhancing the performance of event argument extraction tasks. This method integrates argument extraction related definitions and rule-based knowledge, guiding the extraction process of event arguments through the introduction of heuristic rules, the main prompting process and content are illustrated in Figure 1.

The Argument extraction related definition prompting part mainly focuses on:

Event Attributes and Definitions: Prior to argu-



Model Inputs

Definition-augmented Heuristic-driven Prompting Method:

Argument extraction related definition prompting:

Task definition: Your task is Event Argument Extraction. In this task, you will be provided with a document that describes an event and the goal is to extract the event arguments that correspond to each argument role associated with the event.

The terminologies for this task is as follows:
Event trigger: the main word that most clearly expresses an event occurrence, typically a verb or a noun. The trigger word is located between special tokens "<t>" and "<\t>" in the document, and only the event argument explicitly linked to the trigger word should be considered.
Event argument: an entity mention, temporal expression or value that serves as a participant or attribute with a specific role in an event. Event arguments should be quoted exactly as they appear in the given document.
Argument role: the relationship between an argument to the event in which it participates.[{event_type: Conflict.Attack, trigger: bombing}, {event_type: Life.Die, trigger: killed}]

Argument Extraction Rule Definitions:
-In event argument extraction, arguments are entities or concepts directly related to and participating in events triggered by specific words. They are categorized into roles such as agent, patient, instrument, location, time, cause, and result, and can take the form of named entities, pronouns, noun phrases, verb phrases, adjective phrases, or adverb phrases.
-Each instance may contain one or more events. For each event, it may have one or more argument roles, identify its trigger and list all corresponding arguments found.
-Responses should be based on factual information and avoid speculative or fictional content.

The possible event types and their arguments are as follows:
Life.Die: This event has four arguments (Agent, Victim, Instrument , Place) Agent: The attacking agent / The killer Victim: The person(s) who died, Instrument:The device used to kill Place: Where the death takes place.
Conflict.Attack: This event has four arguments (Attacker, Target, Instrument, Place). Attacker: The attacking/instigating agent, Target: The target of the attack,Instrument: The instrument used in the attack, Place: Where the attack takes placeExample end here.

Example end.

Heuristics-driven CoT:

Heuristics: serving as guiding rules for extracting event arguments.
Specifically, you will use the heuristic provided in the heuristic list to guide identify event arguments, and re-evaluate the identified argument candidates to get the final answer.
heuristic list:
[
Semantic Heuristic: [giver] is the person, group, or organization in the document that gives the grant or gift.
Syntactic Heuristic: The [giver] may be recognized by analyzing sentence structure, often appearing before prepositional phrases starting with 'to' that introduce the recipient (e.g., "X gives Y to Z", X is the 'giver').
Dependency Parsing Heuristic: In parsing the sentence structure, the [giver] is often connected through a dependency relation (e.g., 'nsubj' for nominal subject) to the main verb representing the giving action.
]

CoT:
Step1: Select one or two heuristics in the heuristic list that are most suitable to identify the [argument] in the given document: Semantic Heuristic.
Step2: ...
...

Model Outputs

Figure 1: Definition-augmented Heuristic-driven Prompting method guides on how to extract event arguments related to specific trigger words from documents by defining the task, terminology, extraction rules, and a list of heuristics. It provides corresponding definitions for argument extraction prompting and heuristic rules to assist in the identification and extraction of event arguments.

ment extraction, it is essential to clarify the definition of events and associated terminologies. Events are defined as explicitly marked verbs or nouns in the document, with the verb or noun serving as the event trigger, and event arguments are entities, temporal expressions, or value concepts explicitly connected to this trigger, playing a certain role in the event. For instance, in an event defined as "Conflict.Attack," key event arguments include the attacker (Agent), victim (Victim), weapon (Instrument), and location (Place).

Argument Extraction Rules: We employ a series of heuristic rules to guide the extraction of event arguments. These heuristic rules define potential argument roles based on the relationship between entities and event triggers, such as agents, patients, instruments, locations, times, outcomes, etc., and consider various morphological structures including noun phrases, pronouns, verb phrases, adjective phrases, and adverbial phrases.

The argument extraction related definitions serve as guiding rules for extracting arguments, assisting in the rapid identification of event arguments and reassessing them after identification to determine the final answers. We utilize semantic and dependency parsing heuristics, such as identifying the agent of an action and linking the agent to the verb through dependency relations, to enhance the identification of arguments. Through these definitions, we are able to extract the trigger words for each event and all corresponding arguments, ensuring that the extracted information is fact-based and avoids speculative or fictional content.

For the Heuristics-driven CoT part, we mainly follow the settings and definitions proposed by Zhou et al. (2024) and Wei et al. (2022) to guide the model along a specific logical path, thereby improving the accuracy of event argument extraction. This leverages heuristic rules to inspire and guide the model through a logical chain from preliminary assumptions to final conclusions, revealing the complex structure and associations behind the event. We have optimized parts of the reasoning process:

Initiation Phase: The event triggers and potential arguments identified through Argument extraction related definition prompting initialize the starting point of the reasoning chain. Reasoning Expansion: Based on heuristic rules, the model gradually expands the reasoning chain, parsing the potential relationships and attributes between event arguments through logical deduction. This phase emphasizes adding clear reasoning paths at each step to assist the model in more precise argument extraction in subsequent steps.

Logical Verification: After the reasoning chain is preliminarily constructed, heuristic rules are used to logically verify the reasoning chain, ensuring the rigor of each step and adjusting potential logical errors.

Heuristic rules play a crucial role here, providing a logical foundation and directional guidance for the construction of the Chain-of-Thought. The definitions of these rules are based on an in-depth understanding and recognition of specific event types. For example, by analyzing the linguistic and semantic relationships between event triggers and potential arguments, the logical order and associations of these elements can be deduced.

Through the comprehensive application of these methods, our goal is to enhance the performance of event argument extraction tasks and strengthen the model's ability to grasp the complex relationships between events and arguments.

3 Experiments

3.1 Setup

To evaluate the document-level Event Argument Extraction task, we adopt the RAMS (Ebner et al., 2020) and DocEE (Tong et al., 2022) datasets. For the assessment, we follow the metrics outlined in (Ma et al., 2022; Zhou et al., 2024), which are the F1 score for argument identification (Arg-I) and the F1 score for argument classification (Arg-C). Detailed statistical data of the datasets and the number of test samples are listed in Appendix A. Our Definition-augmented Heuristic-driven Prompting (DHP) method is compared with several state-of-the-art prompting methods, as well as the Chain-of-Thought (CoT) prompting (Wei et al., 2022).

Here, we present the replication of results based on the CoT prompting method by Zhou et al. (2024), which represents one of the few excellent prompting strategies specifically tailored for the Event Argument Extraction task in LLMs. The experiments were conducted using two large language models: the publicly available Deepseek-v2-chat (Liu et al., 2024) and Llama3.1-70b (Dubey et al., 2024). It is noteworthy that due to the relatively high cost of Deepseek-v2-chat, its evaluation was limited to a subset of the dataset. Further experimental details can be found in Appendix A. We also have compared our approach with a variety of

Method		RAMS		DocEE-Normal
		Arg-I	Arg-C	Arg-C
Supervised-learning	EEQA (2020)	-	19.54	-
	PAIE (2022)		29.86	
	TSAR (2022)		26.67	
	CRP (2023)		30.09	
	FewDocAE (2023)		-	12.07
Llama3.1-70b	CoT (2022)	39.80	30.69	26.11
	Ours	**42.33**	**34.60**	**29.69**
Deepseek-v2-chat	CoT (2022)	43.21	38.67	29.67
	Ours	**48.00**	**45.54**	**31.33**

Table 1: Overall performance. In few-shot setting, the scores of supervised learning methods on RAMS dataset are based on results reported in Liu et al. (2023), where 1% of the training data is used.

Method		DocEE-Cross
Supervised-learning	FewDocAE	10.51
Llama3.1-70b	Ours	**32.24**
Deepseek-v2-chat	Ours	**33.43**

Table 2: In the cross-domain setting of the DocEE dataset, the Arg-C performance varies across different methods.

supervised learning methods found in the current literature. These include CRP (Liu et al., 2023), Few-DocAE (Yang et al., 2023), PAIE (Ma et al., 2022), TSAR (Xu et al., 2022), and EEQA (Du and Cardie, 2020). Within the domain of few-shot learning, our comparative analysis is grounded on the performance data from a limited number of samples as previously reported by Liu et al. (2023) and Zhou et al. (2024).

3.2 Results

Table 1 presents experimental results that demonstrate our DHP prompting significantly enhances contextual learning for the document-level Event Argument Extraction (EAE) task.

The DHP method consistently outperforms the CoT prompting (Wei et al., 2022) across LLMs and two datasets. Specifically, in the RAMS dataset, the DHP method achieves the largest F1 score improvements for Arg-I of 2.53% and 4.79%, and for Arg-C of 3.91% and 6.87%, respectively. Compared to supervised learning methods, the application of the DHP method in large models has led to Arg-C score improvements of 4.51% and 15.45%. This indicates that the DHP method significantly enhances the ability of large language models to identify arguments related to specific event triggers and assign them the correct argument roles.

In the DocEE dataset, under normal-setting, our method achieves substantial improvements over FewDocAE, with increases of 17.62% and 19.26%, respectively (Yang et al., 2023). The experimental results suggest that to further ascertain whether the DHP method can enhance the generalization capability of LLMs on data from different domains, which is crucial in real-world applications where large amounts of annotated data may be difficult to obtain (Tong et al., 2022; Luo et al., 2023), we tested the model performance under the Cross domain-settings of the DocEE dataset, as shown in Table 2. The large models with the DHP method also achieved at least a 21.73% increase in the F1 score for Arg-C.

This supports the conclusion that our method can successfully reduce the reliance on large volumes of labeled data for document-level EAE tasks while improving accuracy.

3.3 Analysis

Following our empirical validation of the effectiveness of the DHP method, our approach naturally incorporates various heuristic methods into the prompts. By guiding the model to generate a detailed reasoning process, the accuracy and interpretability of the model are enhanced, which aids in more precisely identifying relationships between entities and improving the accuracy of argument extraction. We decompose the definitions related to the event argument extraction task to avoid performance degradation caused by handling too much information in a single task, thus overcoming the illusion problem. The relevant prompting strategies applied by our DHP method can indeed effectively improve the LLMs performance of unseen classes in the prompts.

We believe that selecting appropriate models and configurations, coupled with carefully designed prompts and balanced datasets, is crucial for improving the performance of event extraction tasks. Moreover, cognitive research has found that compared to complex methods, humans use heuristics as an effective cognitive pathway to achieve more accurate reasoning (Gigerenzer and Gaissmaier, 2011; Hogarth and Karelaia, 2007; Zhou et al., 2024). As similar results presented in the studies by Wei et al. (2022) and Zhou et al. (2024), paralleling this human cognitive strategy, we enable LLMs to learn from explicit heuristics to enhance reasoning. Specifically, for LLMs that perform poorly under vague prompts and in non-reasoning tasks where it is difficult to grasp clear reasons, explicit heuristic specifications provide LLMs with a useful strategy for using and enhancing reasoning. By converting these implicit heuristics into explicit ones, a more direct way to utilize heuristics is provided, allowing LLMs to handle new situations by analogy with known cases. This capability is particularly useful in ICL, as LLMs are always faced with unseen samples and unseen classes (Zhou et al., 2024).

4 Related works

4.1 Document-level EAE

Document-level EAE commonly relies on manual domain and pattern annotation for supervised learning models (Xiang and Wang, 2019; Lin et al., 2020; Li et al., 2022; Liu et al., 2022; Hsu et al., 2022; Liu et al., 2023). The high costs, coupled with the reliance on extensive manually annotated data, may pose a bottleneck for their practical application (Lin et al., 2023). (Agrawal et al., 2022) have employed LLMs in clinical Event Argument Extraction (EAE) using standard prompts that do not involve any reasoning strategies, while research on prompting strategies specifically tailored for the EAE task is scarce, with only (Zhou et al., 2024) exploring the promising and challenging research direction of reducing the dependence on specific large-scale training datasets through ICL, thereby enhancing the generalization capability of LLMs in EAE tasks.

4.2 In-Context Learning

The In-Context Learning (ICL) (Brown et al., 2020) methodology is designed to expedite the adaptability of language models across various tasks, necessitating minimal or no prior data (Wei et al., 2022; Kojima et al., 2022). This methodology eschews direct fine-tuning through the capacity for models to interpret and perform tasks drawing on contextual clues. Weber et al. (2023) enhanced model accuracy by employing carefully crafted efficient prompting templates and diverse prompting formats. Gonen et al. (2023) have noted that the performance of ICL is highly sensitive to the selection of examples. Zhou et al. (2024) innovatively explored the use of examples to guide Large Language Models (LLMs) in processing specific tasks through heuristic rules. This implies that well-designed prompts and heuristic rules can effectively enhance ICL performance without the need for fine-tuning on task-specific datasets.

5 Conclusion

In this study, we propose a Definition-augmented Heuristic-driven prompting strategy for LLMs in document-level event argument extraction tasks. Through experimentation, we have found that incorporating Argument Extraction Related Definition prompting can further enhance the performance of event argument extraction, building upon structured heuristic methods and the Chain-of-Thought approach. Our method has exhibited consistent performance and generalization capabilities across two datasets, showing potential and application prospects.

Limitations

Due to cost constraints, the evaluation of large language models (LLMs) is often limited to a subset of available datasets. This restriction may hinder the comprehensiveness of performance assessments, as a complete dataset could provide a more thorough evaluation, particularly in terms of the advanced reasoning capabilities that LLMs rely on. In this study, we aim to explore the upper limits of contextual learning performance in the EAE task. Our approach leverages the complex reasoning abilities inherent in LLMs, which may not be suitable for models with limited reasoning capabilities. Although we conducted our tests under cross-domain settings using the DocEE dataset, it is important to note that while heuristic rules may perform well on specific tasks and datasets, the generalization capabilities of these models across broader domains and various document types remain an area that warrants further investigation.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual review of psychology*, 62(1):451–482.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148.

Robin M. Hogarth and Natalia Karelaia. 2007. Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, 114(3):733–758.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Degree: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Zizheng Lin, Hongming Zhang, and Yangqiu Song. 2023. Global constraints with prompting for zero-shot event argument classification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2527–2538, Dubrovnik, Croatia. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Jian Liu, Chen Liang, Jinan Xu, Haoyan Liu, and Zhe Zhao. 2023. Document-level event argument extraction with a chain reasoning paradigm. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9570–9583.

Xiao Liu, Heyan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256, Brussels, Belgium. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. 2023. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23103–23123. PMLR.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774.

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.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore. Association for Computational Linguistics.

Meihan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. Docee: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2024. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *arXiv preprint arXiv:2312.04945*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A two-stream AMR-enhanced model for document-level event argument extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5025–5036, Seattle, United States. Association for Computational Linguistics.

Xianjun Yang, Yujie Lu, and Linda Petzold. 2023. Few-shot document-level event argument extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8029–8046.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022a. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Hanzhang Zhou, Junlang Qian, Zijian Feng, Hui Lu, Zixiao Zhu, and Kezhi Mao. 2023. Heuristics-driven link-of-analogy prompting: Enhancing large language models for document-level event argument extraction. *arXiv preprint arXiv:2311.06555*.

Hanzhang Zhou, Junlang Qian, Zijian Feng, Hui Lu, Zixiao Zhu, and Kezhi Mao. 2024. Llms learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. *Preprint*, arXiv:2311.06555.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14638–14646.

A Experimental Details

Dataset	# Example	# Eval.	Eval. Split
RAMS (2020)	1	871	test
DocEE (2022)	1	800	test

Table 3: The overall statistics of the dataset. # Example: The number of examples used in the HDP method. # EVAL.: the number of samples used for evaluation of different prompting methods. EVAL. Split: evaluation split.

The dataset statistics are presented in Table 3. For the large scale of the DocEE and RAMS datasets, full-size evaluation using LLMs is impractical. We follow the setup of Shum et al. (2023); Wang et al. (2022); Zhou et al. (2024), Wang et al. (2022), and Zhou et al. (2024), and evaluate a subset of these datasets. Due to the substantial costs associated with deploying LLMs, we limit our assessment to 200 samples for both the RAMS and DocEE datasets. Furthermore, for the DocEE dataset, it presents two distinct settings. In the conventional configuration, the training and testing data share an identical distribution. Conversely, the cross-domain setup features training and testing data composed of non-overlapping event types (Tong et al., 2022; Zhou et al., 2024).