

SEOE: A Scalable and Reliable Semantic Evaluation Framework for Open Domain Event Detection

Yi-Fan Lu, Xian-Ling Mao*, Tian Lan, Tong Zhang, Yu-Shi Zhu, Heyan Huang

Beijing Institute of Technology

yifanlu@bit.edu.cn, maoxl@bit.edu.cn, lantiangmftby@gmail.com,

tongz@bit.edu.cn, zhuyushi@bit.edu.cn, hhy63@bit.edu.cn

Abstract

Automatic evaluation for Open Domain Event Detection (ODED) is a highly challenging task, because ODED is characterized by a vast diversity of un-constrained output labels from various domains. Nearly all existing evaluation methods for ODED usually first construct evaluation benchmarks with limited labels and domain coverage, and then evaluate ODED methods using metrics based on token-level label matching rules. However, this kind of evaluation framework faces two issues: (1) The limited evaluation benchmarks lack representatives of the real world, making it difficult to accurately reflect the performance of various ODED methods in real-world scenarios; (2) Evaluation metrics based on token-level matching rules fail to capture semantic similarity between predictions and golden labels. To address these two problems above, we propose a scalable and reliable Semantic-level Evaluation framework for Open domain Event detection (SEOE) by constructing a more representative evaluation benchmark and introducing a semantic evaluation metric. Specifically, our proposed framework first constructs a scalable evaluation benchmark that currently includes 564 event types covering 7 major domains, with a cost-effective supplementary annotation strategy to ensure the benchmark’s representativeness. The strategy also allows for the supplement of new event types and domains in the future. Then, the proposed SEOE leverages large language models as automatic evaluation agents to compute a semantic F1-score, incorporating fine-grained definitions of semantically similar labels to enhance the reliability of the evaluation. Extensive experiments validate the representatives of the benchmark and the reliability of the semantic evaluation metric. Existing ODED methods are thoroughly evaluated, and the error patterns of predictions are analyzed, revealing several insightful findings.¹

*Corresponding Author.

¹SEOE is available at <https://github.com/Lyfralston/SEOE>

1 Introduction

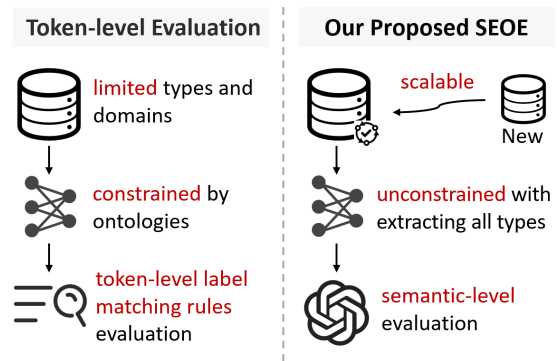


Figure 1: Comparison between previous ODED evaluation frameworks and our proposed SEOE.

Open domain event detection (ODED) is a crucial task in open domain information extraction. Its primary objective is to extract structured information, namely events, from large-scale textual data across various domains without constrained ontologies. This structured information can then be used for further research tasks, such as information retrieval (Zhang et al., 2024; Joshi et al., 2023), knowledge base construction (Zhao et al., 2024), and large language models (Cai et al., 2024; Chen et al., 2024; Wang et al., 2023), where understanding the relationships between events is essential. Recently, the importance of ODED has increased with the rapid expansion of the web, which continuously generates vast amounts of unstructured information across various domains, such as the clinical area (Ma et al., 2023a), suicide understanding (Guzman-Nateras et al., 2022), and scientific publications (Lu et al., 2024b).

Evaluating ODED is a highly challenging task due to the vast diversity of event types in the open domain (Araki and Mitamura, 2018). When making predictions, ODED models are often required to extract events with unseen event types during the

training and validation stage (Cai et al., 2024). To evaluate these models, to the best of our knowledge, nearly all existing ODED evaluation frameworks simulate the open setting by evaluating models using a predefined ontology in a zero-shot manner (Cai et al., 2024; Yue et al., 2023). Specifically, as illustrated in Figure 1, these frameworks typically first construct or utilize evaluation benchmarks with a limited set of types and domain coverage that are unseen during the training stage. During the evaluation process, ODED models are required to recognize events constrained by predefined ontologies of these benchmarks. These evaluation frameworks then apply token-level matching rules to evaluate models’ predictions, such as strict match (Zhang et al., 2022a), partial match (Araki and Mitamura, 2018), and head noun match (Li et al., 2021).

However, the above ODED evaluation frameworks face two issues: (1) The evaluation benchmarks, with limited labels and domain coverage, lack representatives of the real world. Considering such benchmarks as the simulation of the open domain, it is difficult to accurately reflect the performance of ODED models in real-world scenarios; (2) Evaluation metrics based on token-level matching rules fail to capture semantic similarity between predictions and golden event types. Numerous semantically similar event types in open domain cannot be accurately evaluated under these evaluation rules.

To address these two problems above, we propose a scalable and reliable **Semantic Evaluation framework for Open domain Event detection** (SEOE). This framework consists of a more representative evaluation benchmark and a semantic evaluation metric. As for the benchmark, our proposed framework SEOE constructs a scalable evaluation benchmark that currently includes 564 event types covering 7 major domains. We first integrate the ontologies of mainstream event extraction datasets into a larger ontology. Then, we propose a cost-effective supplementary annotation strategy to ensure that each data instance is examined by the integrated ontology, making the benchmark both representative of real-world scenarios and easily scalable. As for the evaluation metric, SEOE leverages large language models (LLMs) as automatic evaluation agents to compute a semantic F1-score. To ensure reliable evaluation, SEOE generates fine-grained definitions of event types and groups semantically similar labels together to assist in the

evaluation process.

As for the experiments, the representatives of the evaluation benchmark and the reliability of the semantic evaluation metric are validated. With such a representative benchmark and a reliable metric in SEOE, we conduct comprehensive evaluation experiments on existing ODED methods including closed-source and open-source LLMs. Extensive experiments reveal that ODED remains a highly challenging task, particularly in balancing the accuracy and diversity of the predictions. Furthermore, a systematic analysis about the patterns of the model’s incorrect predictions reveals that predicting event types and generating their definitions in the ODED task is a highly challenging problem.

In summary, the contributions of our study are two-fold:

- **SEOE:** We propose a scalable and reliable semantic evaluation framework for ODED, which is built upon a more representative evaluation benchmark and a semantic evaluation metric, with LLM serving as the automatic evaluation agent. SEOE achieves cost-effective scalability and demonstrates strong correlations with human evaluations.
- **Comprehensive evaluation and analysis:** We apply our proposed SEOE to comprehensively evaluate the performance of existing ODED models. We systematically analyze the patterns of the models’ incorrect predictions. The experiments reveal insightful phenomenon worth further investigation.

2 Related Works

2.1 Event Detection

Event detection (ED), as an important sub-task of event extraction in information extraction, can be divided into closed-domain ED and open-domain ED based on the presence of a predefined event schema (Liu et al., 2021a). Due to the increasing number of domains of interest (Sun et al., 2022) and expensive expert annotations when constructing closed-domain datasets (Hsu et al., 2022), ODED research has gained significant attention (Veyseh et al., 2021; Liu et al., 2019). Besides, some studies on few-shot (Ma et al., 2023b) and zero-shot ED (Yue et al., 2023; Zhang et al., 2022b) can emulate the open-domain setting. However, these studies provide predefined event types during model prediction, which does not align with real-world open-domain scenarios.

2.2 LLMs for Automatic Benchmark Construction

With the emergence of advanced LLMs, exemplified by GPT-4, automated data annotation and synthesis have garnered significant attention in the NLP community (Tan et al., 2024). Many studies on LLM-based annotation focus on automating the labeling process with the help of LLMs (Yadav et al., 2024; Tseng et al., 2024), including single LLM annotation (Chen et al., 2024; Xu et al., 2025), joint annotation by multiple LLMs (Martorana et al., 2024), and hybrid annotation by humans and LLMs (Li et al., 2023). In addition, the evaluation of annotations is also a crucial part of automated annotation, with mainstream studies including general evaluation with human comparisons (Honovich et al., 2023), task-specific evaluations (Chen et al., 2023b), and using LLMs as evaluators known as LLM-as-a-judge (Wu et al., 2024; Zheng et al., 2023). In this paper, we propose a cost-effective method to instruct GPT-4o to construct a representative ODED benchmark.

2.3 Evaluation for Event Detection

ED is typically evaluated using token-level matching rules (Li et al., 2021; Lin et al., 2020), which assess whether the model-generated events and their labels are token-level identical or similar to the human-annotated ground truth. However, these rules exhibit weak correlation with human judgments due to the lack of the semantic evaluation (Lu et al., 2024a). Recently, there is an emerging trend in the natural language generation research community to leverage LLMs for evaluating generations in a semantic manner (Li et al., 2025; Lan et al., 2024a; Ma et al., 2025; Tu et al., 2024). Notably, (Lu et al., 2024a) propose a LLM-based automatic evaluation framework for event extraction, named RAEE. However, RAEE is designed for closed-domain scenarios and faces the challenge of evaluating a wide range of event types with varying definitions in open-domain scenarios. In this paper, we propose a LLM-based automatic evaluation framework that includes a scalable benchmark and a reliable evaluation metric to semantically evaluate ODED models.

3 ODED Task Formulation

To the best of our knowledge, nearly all existing ODED evaluation frameworks leverage limited evaluation benchmarks to evaluate ODED models

with constrained event types in a zero-shot way. This kind of evaluation framework cannot reflect ODED methods’ performance in real-world scenarios. Therefore, it is necessary to formally define the ODED task for further evaluation.

Given a text sequence $T = \{w_1, w_2, \dots, w_k\}$, where w_i represents i -th word in the text, ODED models are tasked with identifying event triggers $\mathcal{E} = \{e_1, e_2, \dots, e_m\}$. Each event trigger mention comprises words from T . In addition to identifying event trigger mentions, models are required to predict event types, denoted as \mathcal{C} , for each trigger, without relying on pre-defined ontologies. Furthermore, models must provide fine-grained definitions $\mathcal{D} = \{(t, d) \mid t \in \mathcal{C}\}$ for each event type.

4 Open Domain Event Benchmark

Nearly all existing ODED evaluation benchmarks are limited in the diversity of event types and domains, which are not representative of real-world scenarios. To solve this problem, we propose a scalable and more representative benchmark. We will first introduce the construction process in Section 4.1 and then analyze it in Section 4.2 and 4.3.

4.1 Benchmark Construction

For a comprehensive benchmark, it is widely agreed that each data instance should be examined across all predefined types, regardless of whether these types are ultimately annotated (Zhang and Zhou, 2013). However, this requirement is extremely costly for an ODED evaluation benchmark, given the large number of event types across various domains and the potential scalability of the benchmark in terms of event types. To solve this issue, we leverage GPT-4o² to enable automatic annotation and propose a cost-effective method to ensure that each data instance is examined by all event types from an integrated ontology. Figure 2 illustrates the process of benchmark construction in the following four steps. Related prompts can be found in Appendix E.1.

Step 1: Ontology and Data Integration It is important for an ODED dataset to include a diversity of event types and a breadth of domains. Similar to UniversalNER (Zhou et al., 2023), we integrate mainstream event datasets’ ontologies into a large ontology and sample a subset of the data by uniformly sampling cases across all types. More details about the integration is shown in Appendix A.

²<https://openai.com/index/hello-gpt-4o/>

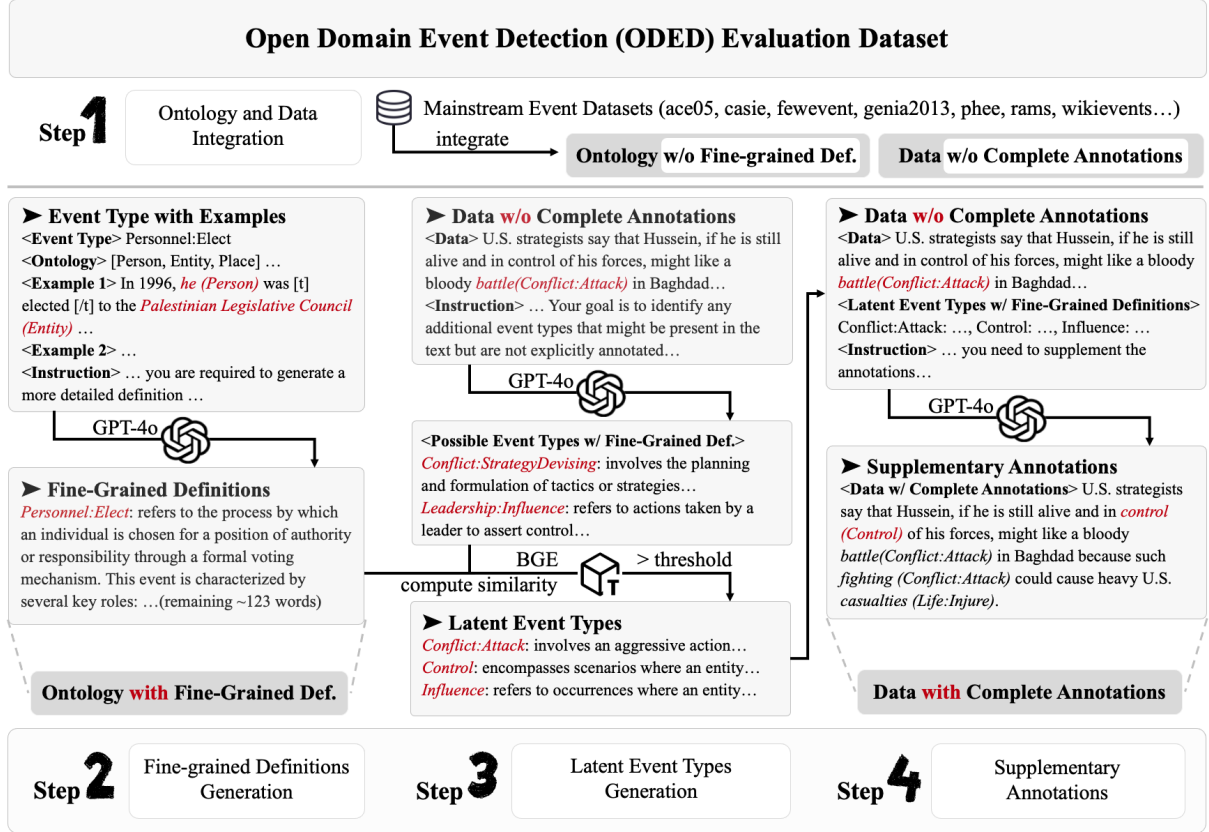


Figure 2: The pipelined construction process of proposed open domain event detection evaluation dataset.

An overview of data distribution is shown in Figure 3. However, this kind of integration introduces an issue that data from one dataset is not examined by the ontologies of other datasets. We will address it in steps 3 and 4.

Step 2: Fine-grained Definitions Generation

Fine-grained definitions of event types help models better understand the event types themselves (Cai et al., 2024). It is crucial to consider these definitions during evaluation, rather than relying solely on their names. Therefore, after integrating a broader ontology, we instruct GPT-4o to generate a fine-grained definition for each event type based on its event ontology and several examples. These generated definitions average approximately 109 words, including a detailed description of the event type, an interpretation of its roles, and specific requirements derived from the examples.

Step 3: Latent Event Types Recognition To solve the problem that a data instance is not examined by the entire ontology, we propose a cost-effective method to achieve the complete annotation in the following two steps. Specifically, in this step, we first instruct GPT-4o to identify possible

event types and generate their corresponding fine-grained definitions. Then, a text similarity model (Chen et al., 2023a) is employed to compute pairwise similarities between possible event types and the integrated event types. Integrated event types that either rank in the top- k for similarity with possible event types, or exceed a predefined similarity threshold, are considered as latent event types and will be used in the next step. In our implementation, we set $k = 5$ and the threshold as 0.8. Through this step, we filter out irrelevant event types from the text, thereby significantly reducing annotation costs. A more detailed formal discussion of the cost can be found in Appendix D.1.

Step 4: Supplementary Annotations For latent event types that may be present in the text but are not explicitly annotated, we provide their fine-grained definitions along with the incomplete data to GPT-4o for supplementary annotations. A full implementation example is illustrated in Appendix E.1. After this step, each data instance in the benchmark is completely annotated, which aligns with the characteristic of real-world open domain scenarios where each data should be examined by all possible labels.

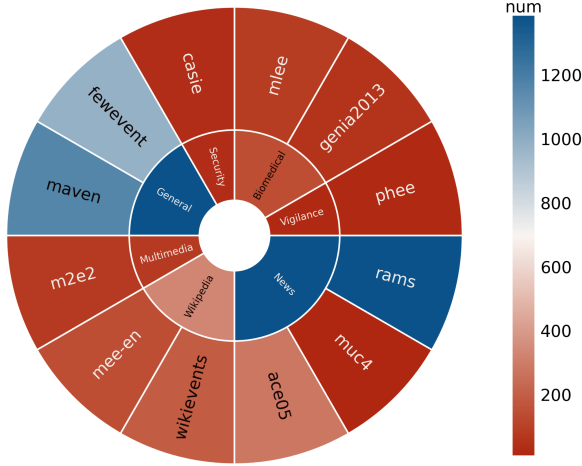


Figure 3: Distribution of SEOE benchmark.

4.2 Benchmark Reliability Enhanced by Nucleus Sampling

A benchmark based on LLMs generation may suffer from issues related to the reliability of annotations (Chen et al., 2024; Tan et al., 2024). In the construction steps outlined above, the reliability of our benchmark is mainly influenced by the **robustness** and **quality** of the single LLM-generated annotation. Therefore, it is crucial to introduce repeated annotations and a sampling strategy to ensure the reliability of our benchmark.

Inspired by nucleus sampling in text generation (Holtzman et al., 2019), we repeat the LLM annotation process in Figure 2, and calculate the frequency of each trigger across multiple annotation rounds. These triggers are then sorted by frequency in descending order, and those with a cumulative frequency reaching a threshold p are selected. Next, we instruct GPT-4o to merge semantically redundant triggers. With this annotation strategy, we enable adjustable sampling by controlling the number of annotation rounds and the threshold p .

Annotation Strategy	Trg. Count	Correct Trg. (%)
1 round	310	80.43 _{1.59}
3 rounds w/. $p = 0.5$	362	81.12 _{1.88}
5 rounds w/. $p = 0.5$	389	83.63 _{3.65}
10 rounds w/. $p = 0.3$	290	86.90 _{2.69}
10 rounds w/. $p = 0.5$	404	85.40 _{2.46}
10 rounds w/. $p = 0.7$	522	81.80 _{4.07}

Table 1: The results of different annotation strategies on the number of supplementary triggers (*Trg. Count*) and annotation accuracy (*Correct Trg.*). There is a trade-off between the diversity and accuracy of the supplementary annotations.

To validate the above strategy’s effectiveness and the benchmark’s reliability, we sample 200 documents (2,277 events in total) and conduct a quality check by 3 humans on the supplementary annotation data, as shown in Table 1. Results show that (1) With a fixed p , increasing annotation rounds improves both the number and accuracy of supplementary triggers. This demonstrates that multiple annotations and nucleus sampling strategy are **highly effective** in enhancing the benchmark’s reliability. (2) When the number of annotation rounds is fixed, an increase in p leads to a rise in the total number of supplementary annotations, but the annotation accuracy decreases. This suggests that adjusting p represents a **trade-off between the diversity and accuracy** of the supplementary annotations. A more detailed analysis can be found in Appendix C. We will release three versions of the evaluation benchmark, based on 10 rounds of repeated annotations and using $p = 0.3$, $p = 0.5$, and $p = 0.7$, allowing future research to choose an evaluation that favors accuracy ($p = 0.3$) or diversity ($p = 0.7$). In subsequent experiments, we will consistently use the balanced version ($p = 0.5$) for evaluation.

4.3 Benchmark Statistics and Scalability

After multiple rounds of supplementary annotation and nucleus sampling, the number of events in the benchmark increases by a factor of 2.29 compared to the original data. More detailed statistics can be found in the Appendix B.

Since ODED aims to detect all possible events from any potential domain, it is essential to ensure the scalability of the ODED evaluation benchmark. In SEOE, both data (i.e., event-containing documents) and event types are scalable. Specifically, when new data is added to the benchmark, step 3 and 4 are repeated to achieve new annotations. When new event types are added, step 2 generates their fine-grained definitions. Then, step 3 and 4 perform new supplementary annotations for previous data. Since step 3 employs a text similarity model for filtering, the additional cost incurred from adding new data and event types is approximately linear with the number of data instances and event types. This ensures the benchmark’s cost-effective scalability and enhances the model’s representativeness of real-world scenarios. To support this conclusion, a formulaic analysis can be found in Appendix D.2.

5 Semantic Evaluation by LLMs

With a clear task definition and a representative evaluation benchmark, we will introduce how to achieve semantic evaluation for ODED. Section 5.1 includes a proposed module to assist in evaluation. Section 5.2 includes how to apply LLMs as automatic evaluation agents to compute a semantic F1-score. Finally, we will show that GPT-4o’s evaluation has a strong correlation with human judgments in Section 5.3.

5.1 Semantically Similar Fine-grained Definitions as Groups

Previous work validates that ontology information, such as relationships among event types, helps models with the understanding of the event, and generalize better to unseen event types (Cai et al., 2024). Since our integrated ontology lacks such information, we propose a definition-aware method to construct it. Specifically, we leverage fine-grained definitions generated from Section 4.1 to compute the similarity between each pair of event types by applying an off-the-shelf text similarity model (Chen et al., 2023a). Event type pairs with similarity exceeding a predefined threshold are considered highly related, and all such related event types are regarded as a **group**, as illustrated in Figure 4. In the following evaluation process, we will provide all event types within the group of the evaluated event type as the ontology information. The effectiveness of such group module will be validated in Section 5.3.

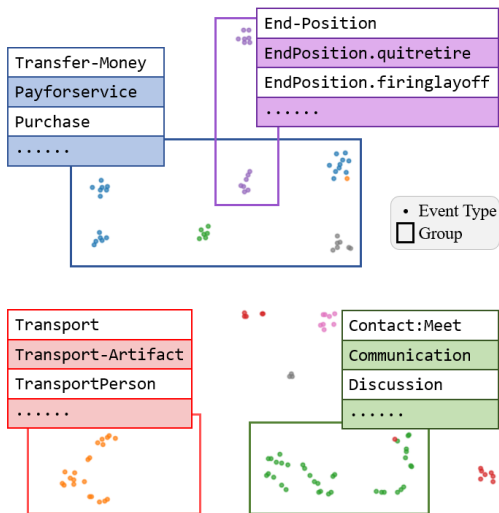


Figure 4: T-SNE (van der Maaten and Hinton, 2008) visualization results of our constructed groups. To maintain conciseness, we only present a subset of groups.

5.2 Semantic Evaluation Process

We first formalize the evaluation task for ODED. Suggest there is a tuple $\{T, \mathbf{P}, \mathbf{D}_p, \mathbf{G}, \mathbf{D}_g\}$, where T is an extracted text, \mathbf{P} is a predicted trigger set from an open ED model, \mathbf{D}_p is the predicted event definition set, \mathbf{G} is a golden trigger set from a benchmark, and \mathbf{D}_g is the golden event definition set. An ODED evaluation framework is required to identify the correctness of each prediction in \mathbf{P} and the recall of each annotation in \mathbf{G} .

In SEOE, similar to (Lu et al., 2024a), we utilize LLMs as the automatic semantic evaluation agent. The above tuple along with corresponding group information as mentioned in Section 5.1 will be prompted to the agent. The agent is required to identify a semantic correspondence set \mathbf{C} . The elements in \mathbf{C} , denoted as (p, g) , where $p \in \mathbf{P}, g \in \mathbf{G}$, indicate that prediction p and golden annotation g are semantically matched. Then, semantic-level F1 score can be computed as follows:

$$\begin{aligned} \mathbf{C}_p &= \{p \mid (p, g) \in \mathbf{C}\}, \\ \mathbf{C}_g &= \{g \mid (p, g) \in \mathbf{C}\}, \\ \mathbf{p} &= \frac{|\mathbf{C}_p|}{|\mathbf{P}|}, \quad \mathbf{r} = \frac{|\mathbf{C}_g|}{|\mathbf{G}|}, \quad \mathbf{F1} = \frac{2 * \mathbf{p} * \mathbf{r}}{\mathbf{p} + \mathbf{r}} \end{aligned}$$

5.3 LLMs’ Evaluation Correlation with Human Judgments

To prove the effectiveness of our proposed SEOE framework and several modules, we conduct a meta-evaluation to analyze the correlation between SEOE and human judgments. Specifically, we invite 3 professional event annotators to evaluate 200 event-containing documents (791 prediction-gold event pairs in total). These annotators are given the same prompts as those given to LLMs. As for the evaluation agents, we select the top-3 LLMs that demonstrate superior performance in text evaluation³. As for the evaluation metrics, we compute three widely-used inter-annotator agreement (IAA) metrics among humans and LLMs: percent agreement (Zheng et al., 2023; Liu et al., 2021b), Spearman correlation (Lan et al., 2024b), and Cohen’s Kappa (Thakur et al., 2024). To minimize the impact of randomness from LLMs, we repeat the experiment three times and report the average and standard deviation of the results, as shown in Table 2.

According to Table 2, it is observed that GPT-4o outperforms other LLMs in terms of IAA with

³<https://openai.com/index/hello-gpt-4o/>

Comparison	Percent	Spearman	Cohen's
3 Humans	95.32	79.92	79.54
GPT-4o	94.41 _{0.14}	77.50 _{0.39}	77.03 _{0.44}
- w/o. Groups	93.37 _{0.20}	74.63 _{0.52}	73.73 _{0.61}
- w/o. #Defs	92.97 _{0.28}	73.85 _{0.94}	72.67 _{1.00}
Claude3-Opus	93.95 _{0.03}	76.03 _{0.35}	75.43 _{0.27}
GPT-4-turbo	93.78 _{0.09}	74.99 _{0.54}	74.50 _{0.49}

Table 2: Three inter-annotator agreement (IAA) metrics (%) among humans and LLMs. *#Defs*: Fine-grained Definitions. *3 Humans* reports the IAA among humans, regarded as the golden standard. IAA between a LLM and each human evaluator is calculated first, and then the average value is taken. All Spearman correlation scores are associated with the p -value < 0.005 .

human evaluation. Its evaluation results are highly aligned with pure human assessments, with a percent agreement difference of 0.91%, a Spearman correlation difference of 2.42%, and a Cohen’s Kappa difference of 2.51%. Furthermore, the introduction of group information and fine-grained definitions significantly improves IAA among GPT-4o and humans. We will use GPT-4o as the automatic evaluation agent in the subsequent evaluation experiments.

6 Evaluation Experiments

With a complete and widely domain-covered benchmark along with a reliable evaluation process, we will evaluate existing ODED models in this section. Apart from the main evaluation results in Section 6.2, we will analyze patterns of incorrect predictions of evaluated ODED models in Section 6.3.

6.1 Experimental Settings

Following the ODED task formulation in Section 3, it is challenging for small language models to identify cases of unseen event types (Veyseh et al., 2021) along with fine-grained definitions. Therefore, we only conduct evaluation experiments on existing closed-source and open-source LLMs.

To mitigate the impact of randomness caused by the stochastic process in LLMs (Bouras, 2024), we will use deterministic decoding by setting the temperature parameter of all evaluated LLMs to 0.

6.2 Overall Analysis of Evaluation Results

Table 3 shows LLMs’ ODED results evaluated by SEOE, indicating several findings: (1) Closed-source LLMs slightly outperform open-source

Models	Trigger Classification (%)		
	precision	recall	F1-score
<i>Closed-source GPT Series LLMs</i>			
o1-preview	62.67	56.59	59.47
GPT-4	67.46	45.12	54.08
GPT-4-turbo	66.53	45.23	53.85
GPT-4o	73.58	45.61	56.31
GPT-3.5-turbo	69.43	42.89	53.02
<i>Closed-source Claude Series LLMs</i>			
Claude-3.5-haiku	75.35	44.26	55.76
Claude-3.5-sonnet	68.47	48.20	56.57
<i>Closed-source Gemini Series LLMs</i>			
Gemini-1.5-flash	64.68	52.41	57.90
Gemini-1.5-pro	63.45	49.19	55.42
<i>Open-source Qwen Series LLMs (Bai et al., 2024)</i>			
Qwen-2.5-72B-Chat	68.83	49.20	57.38
Qwen-2.5-7B-Chat	71.16	43.43	53.94
<i>Open-source InternLM Series LLMs (Team, 2023)</i>			
InternLM-2.5-20B	72.08	43.84	54.52
InternLM-2.5-7B	71.64	42.88	53.65
<i>Open-source Llama-3 Series LLMs (Dubey et al., 2024)</i>			
Llama-3.2-70B-Chat	72.28	45.77	56.05
Llama-3.1-8B-Chat	74.77	42.73	54.38

Table 3: Closed-source and open-source LLMs’ open domain event detection results evaluated by SEOE.

LLMs on the ODED task. Among the closed-source models, o1-preview achieves the best performance with an F1-score of 59.47. The best-performing open-source model is Qwen-2.5-72B, with an F1-score of 57.38. (2) The results from open-source models reveal a clear trend: as the number of model parameters increases, the model’s F1-score improves, with this improvement primarily driven by gains in recall. (3) Across all LLMs, the F1-scores are mainly affected by relatively low recall rather than precision, indicating that while LLMs ensure the accuracy of extracted event triggers, they may lack the capability to comprehensively identify a broader range of triggers.

To further validate statement (3), we select the top-3 LLMs with the highest recall scores and the bottom-3 LLMs with the lowest recall scores from Table 3, and calculate their average number of predicted triggers, as shown in Table 4. The results indicate a positive correlation between recall and the average number of predicted triggers, and a negative correlation with precision. It further reveals a trade-off between **accuracy** and **diversity** in model predictions for the ODED task. Besides, this phenomenon is highly similar to the findings

discussed in Section 4.2, suggesting that ODED models may improve the quality of open domain extraction results through multiple extractions with high-quality sampling (Holtzman et al., 2019).

Models	precision	recall	#TrgNum
o1-preview	62.67	56.58	4.59
Gemini-1.5-flash	64.68	52.41	4.17
Qwen-2.5-72B-Chat	68.83	49.20	3.59
15 LLMs in Average	—	—	3.39
GPT-3.5-turbo	69.43	42.89	3.02
InternLM-2.5-7B	71.64	42.88	2.81
Llama-3.1-8B-Chat	74.77	42.73	2.74

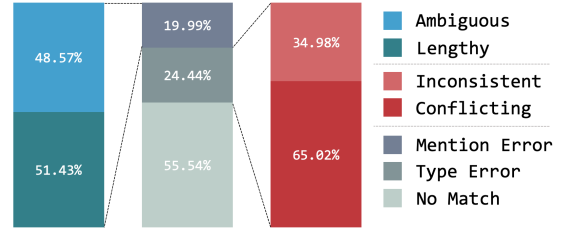
Table 4: Top-3 and bottom-3 recall score LLMs’ average number of predicted triggers (#TrgNum).

6.3 Incorrect Predictions Analysis

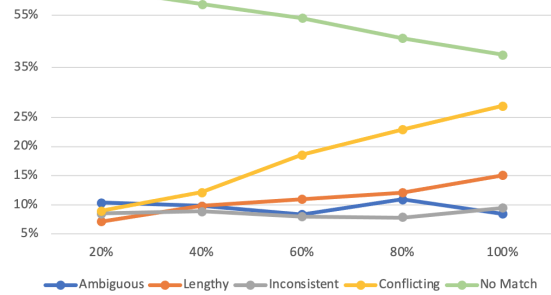
The main characteristic of ODED is that models are not restricted to predefined event types during extraction, and are required to generate event types with definitions. Therefore, the error classification in closed domain (Chen et al., 2024; Lu et al., 2024a) is not applicable to analyze incorrect predictions in ODED. Based on the observation of human quality check in Section 4.2, we ascribe incorrect predictions to several patterns, including:

- **Ambiguous Mention:** The mention of prediction is too brief, leading to unclear semantics, making it difficult to determine the event it refers to.
- **Lengthy Mention:** The mention of prediction is too lengthy, complicating the accurate identification of a certain event.
- **Inconsistent Type Definition:** With a semantic golden match to the mention, the definition of predicted type is overly broad or too narrow compared to that of the golden match.
- **Conflicting Type Definition:** With a semantic golden match to the mention, the definition of predicted type has conflicts with the definition of the golden match.
- **Reasonable Prediction with No Match:** The prediction can be considered reasonable but lacks a semantic golden match.

We manually annotate the distribution of the above error patterns of the incorrect predictions across the entire benchmark, as shown in Figure 5.



(a) Overall error pattern distribution.



(b) The trend of error distribution with the variation in event density across texts. The event density is defined as the number of events in a text divided by the text length. We sort the data based on event density and then present error distribution at different cumulative rankings.

Figure 5: Error Pattern Distribution. All data are randomly sampled from the main experiment.

Overall, the most frequent error type is *No Match* (55.54%), and its proportion decreases as the event density of the text increases. This phenomenon can be interpreted in two aspects: (1) it suggests that the ODED benchmark may be lacking certain event types, highlighting the need for the benchmark’s scalability; (2) since the benchmark’s ontology is derived from mainstream event extraction research, it indicates that a significant proportion of events extracted by the evaluated ODED models are not covered by existing event extraction studies, further suggesting that the extraction strategy of ODED models needs targeted adjustments.

Additionally, overall, the proportion of errors in the *type* category (24.44%) exceeds that in the *mention* category (19.99%), and this gap is even more pronounced in documents with higher event density. It indicates that compared to extracting mentions, determining the type of a mention and defining the type is a more challenging task in ODED.

7 Conclusion and Future Works

In this paper, we propose a scalable and reliable ODED evaluation framework, named SEOE. It consists of a more representative benchmark and a well-validated semantic evaluation process. Experiments reveal that ODED is a challenging task, espe-

cially in predicting event types and corresponding definitions. In the future, we will explore the evaluation of open domain event argument extraction and investigate methods to enhance open domain event extraction.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (No. 62172039, 6240074151, U21B2009 and 62276110).

Limitations

LLMs are not Perfect Human Substitute In this paper, advanced LLMs are used as evaluation agents, showing very strong correlation with human judgments. However, it is important to note that LLMs cannot perfectly align with human evaluations. In the future, we would like discover other strategies to further improve the reliability of LLM-as-a-Judge, such as multi-agent debate (Chan et al., 2023).

Correctness of Generated Definitions In our annotation and evaluation experiments, GPT-4o is used to generate fine-grained definitions for all event types to assist with supplementary annotation and semantic evaluation. While experiments have demonstrated the effectiveness of this module, it is important to note that these fine-grained definitions might have errors.

Cost of LLM-based Evaluation Unlike evaluations based on token-level matching rules, our evaluation framework utilizes advanced LLMs to assess models’ ODED predictions, which introduces higher costs. In the case of the full benchmark, which contains 4,423 documents, the cost of evaluating a model is approximately \$59.8. However, it is possible to accurately evaluate models using only a subset of the benchmark, as the average frequency of each event type in our benchmark is approximately 50.8, which may exceed the requirements for model evaluation. To validate this hypothesis, we identify subsets of the benchmark with sizes of 1/10, 1/4, and 1/2 through random sampling. The average differences in evaluation metrics between these subsets and the full benchmark used in the main experiments are 1.06%, 0.57%, and 0.38%, respectively. This indicates that using an appropriately chosen subset can significantly reduce evaluation costs while maintaining statistical

representativeness. These subsets of the benchmark will also be publicly available.

Evaluation without Offsets In our proposed evaluation framework, we do not apply offsets to the golden annotations in the benchmark, nor do we require the evaluated models to output the corresponding offsets for their predictions. This setting might introduce ambiguity when events with the same mention and event type appear at different positions in the text. Our evaluation decision is based on three main considerations: (1) Adding offsets to the generated results is a common challenge in generative extraction tasks. Some studies (Lu et al., 2022) have shown that this issue occurs in a small proportion of cases and can be addressed using simple heuristics, such as mapping spans to offsets by finding the first matched offsets, which results in minimal errors (<0.5%). (2) The ability of ODED models to output offsets is not sufficiently reliable, and requiring them to do so could lead to less accurate evaluation results. (3) The proposed semantic-level evaluation does not depend on tokens, so the need for offsets is minimal. We plan to explore the specific impact of offsets on ODED evaluation in future work.

References

- 1992. *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.
- Jun Araki and Teruko Mitamura. 2018. [Open-domain event detection using distant supervision](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.
- Andrew Bouras. 2024. [Integrating randomness in large language models: A linear congruential generator approach for generating clinically relevant content](#). Preprint, arXiv:2407.03582.
- Zefan Cai, Po-Nien Kung, Ashima Suvarna, Mingyu Ma, Hritik Bansal, Baobao Chang, P. Jeffrey Brantingham, Wei Wang, and Nanyun Peng. 2024. [Improving event definition following for zero-shot event detection](#). In *Proceedings of the 62nd Annual Meeting*

- of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2842–2863, Bangkok, Thailand. Association for Computational Linguistics.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. [Chateval: Towards better llm-based evaluators through multi-agent debate](#). *Preprint*, arXiv:2308.07201.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023a. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Preprint*, arXiv:2309.07597.
- Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. [Is a large language model a good annotator for event extraction?](#) *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17772–17780.
- Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023b. [DISCO: Distilling counterfactuals with large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.
- Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM)*.
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. [The automatic content extraction \(ACE\) program – tasks, data, and evaluation](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- 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*, pages 8057–8077, Online. Association for Computational Linguistics.
- Luis Guzman-Nateras, Viet Lai, Amir Poursan Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. 2022. [Event detection for suicide understanding](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1952–1961, Seattle, United States. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. [Unnatural instructions: Tuning language models with \(almost\) no human labor](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.
- 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, Seattle, United States. Association for Computational Linguistics.
- Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. [TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12804–12825, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. [U-CREAT: Unsupervised case retrieval using events extrAcTion](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13899–13915, Toronto, Canada. Association for Computational Linguistics.
- Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. [The Genia event extraction shared task, 2013 edition - overview](#). In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria. Association for Computational Linguistics.
- Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. 2024a. [Training language models to critique with multi-agent feedback](#). *Preprint*, arXiv:2410.15287.
- Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-ling Mao. 2024b. Criticbench: Evaluating large language models as critic. *arXiv preprint arXiv:2402.13764*.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From generation to judgment: Opportunities and challenges of llm-as-a-judge](#). *Preprint*, arXiv:2411.16594.
- Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020.

- Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.
- Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. [CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.
- Sha Li, Heng Ji, and Jiawei Han. 2021. [Document-level event argument extraction by conditional generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.
- 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*, pages 7999–8009, Online. Association for Computational Linguistics.
- Jiangwei Liu, Liangyu Min, and Xiaohong Huang. 2021a. [An overview of event extraction and its applications](#). *Preprint*, arXiv:2111.03212.
- Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021b. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13415–13423.
- Xiao Liu, Heyan Huang, and Yue Zhang. 2019. [Open domain event extraction using neural latent variable models](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871, Florence, Italy. Association for Computational Linguistics.
- Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. [Unified structure generation for universal information extraction](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.
- Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Chen Xu, and Heyan Huang. 2024a. [Beyond exact match: Semantically reassessing event extraction by large language models](#). *Preprint*, arXiv:2410.09418.
- Yi-Fan Lu, Xian-Ling Mao, Bo Wang, Xiao Liu, and Heyan Huang. 2024b. [Exceeds: Extracting complex events as connecting the dots to graphs in scientific domain](#). *Preprint*, arXiv:2406.14075.
- Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023a. [DICE: Data-efficient clinical event extraction with generative models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. Association for Computational Linguistics.
- Yubo Ma, Zehao Wang, Yixin Cao, and Aixin Sun. 2023b. [Few-shot event detection: An empirical study and a unified view](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11211–11236, Toronto, Canada. Association for Computational Linguistics.
- Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Yu-Shi Zhu, Tong Zhang, Heyan Huang, and Xian-Ling Mao. 2025. [Multi-modal retrieval augmented multi-modal generation: Datasets, evaluation metrics and strong baselines](#). *Preprint*, arXiv:2411.16365.
- Margherita Martorana, Tobias Kuhn, Lise Stork, and Jacco van Ossensbruggen. 2024. Zero-shot topic classification of column headers: Leveraging llms for metadata enrichment. In *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI*, pages 52–66. IOS Press.
- Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. [MEE: A novel multilingual event extraction dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Junichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):575–581.
- Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.
- Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. [PHEE: A dataset for pharmacovigilance event extraction from text](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

- InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.
- Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2024. Are expert-level language models expert-level annotators? *Preprint*, arXiv:2410.03254.
- Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, Heyan Huang, and Xian-Ling Mao. 2024. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. *Preprint*, arXiv:2411.15488.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.
- Amir Pouran Ben Veyseh, Minh Van Nguyen, Bonan Min, and Thien Huu Nguyen. 2021. Augmenting open-domain event detection with synthetic data from gpt-2. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 644–660. Springer.
- Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansan Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv:2304.08085.
- Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.
- Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *Preprint*, arXiv:2407.19594.
- Chen Xu, Tian Lan, Yu Ji, Changlong Yu, Wei Wang, Jun Gao, Qunxi Dong, Kun Qian, Piji Li, Wei Bi, and Bin Hu. 2025. Decider: A dual-system rule-controllable decoding framework for language generation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14.
- Sachin Yadav, Tejaswi Chopra, and Dominik Schlechtweg. 2024. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4. *Preprint*, arXiv:2407.04130.
- Zhenrui Yue, Huimin Zeng, Mengfei Lan, Heng Ji, and Dong Wang. 2023. Zero- and few-shot event detection via prompt-based meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7928–7943, Toronto, Canada. Association for Computational Linguistics.
- Hongming Zhang, Wenlin Yao, and Dong Yu. 2022a. Efficient zero-shot event extraction with context-definition alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7169–7179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.
- Senhui Zhang, Tao Ji, Wendi Ji, and Xiaoling Wang. 2022b. Zero-shot event detection based on ordered contrastive learning and prompt-based prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2572–2580, Seattle, United States. Association for Computational Linguistics.
- Yanan Zhang, Xiaoling Bai, and Tianhua Zhou. 2024. Event-enhanced retrieval in real-time search. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6594–6606, Torino, Italia. ELRA and ICCL.
- Bang Zhao, Yilong Zhao, and Ying Mao. 2024. A method for judicial case knowledge graph construction based on event extraction. In *Proceedings of the 2024 9th International Conference on Intelligent Information Technology, ICIIT '24*, page 62–69, New York, NY, USA. Association for Computing Machinery.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.

A Benchmark Integration

We utilize a unified event extraction framework (Huang et al., 2024) to perform the integration. Basic statistics about sampled datasets are presented in Table 5.

During the integration, we uniformly sample cases across all types. Specifically, during the sampling process, we try to ensure that each type ap-

pears in at least 10 documents. Once a type appears in 10 documents, we stop sampling it further.

We do not merge event types with the same name from different datasets originating from the same ontology. This is because different datasets may have their own customized annotation or interpretation rules for the same ontology, which are inflected in their distinctive annotation guidelines. We add serial numbers to distinguish event types with the same name. This treatment also enhances the feasibility of scaling our benchmark.

B Benchmark Statistics

Table 6 presents basic statistics of 3 different versions of our proposed SEOE’s benchmark. The benchmark with $p = 0.3$ focuses more on evaluating the accuracy of the model’s extraction results, while the benchmark with $p = 0.7$ emphasizes the evaluation of the model’s extraction diversity. The benchmark with $p = 0.5$ is a balanced version and is used in the evaluation experiments presented in the main text.

We also provide a comparison of event type distribution in SEOE and other datasets, shown in Table 7.

C Benchmark Reliability Analysis

In the main text, we report the annotation accuracy under different annotation strategies, using only the classifications of *correct* or *incorrect*. In fact, during the annotation process, we further subdivide these two types into the following seven subtypes:

- C1: Nearly perfect annotations.
- C2: Semantically correct, but the mention is slightly longer than necessary.
- C3: Semantically correct, but the event type is slightly too broad.
- C4: Semantically correct, but semantically redundant with one of the golden annotations.
- IC1: Incorrect annotation due to an incorrect mention.
- IC2: Incorrect annotation due to an incorrect event type.
- IC3: Completely incorrect annotation at both the mention and event type levels, or the selection of words not present in the original text.

The proportion of the above 7 subtypes under different annotation strategies is illustrated in Table 8 and 9. It is worth noting that, although we use GPT-4o to merge semantically redundant annotations after the nucleus sampling, as mentioned in the main text, a substantial proportion of semantically redundant annotations still remain (i.e., C4 in Table 8). Further observation reveals that this redundancy is primarily caused by two main factors: (1) similar events occurring at different positions in the text, and (2) different perspectives on the same event, such as the text *A mother gave birth to a baby*, which might involve both a *Life: Pregnancy* event and a *Life: Born* event, leading to semantic overlap. Nevertheless, these two factors do not affect the accuracy of the subsequent evaluation, and therefore we still consider these annotations as correct.

D Benchmark Cost and Scalability Analysis

Our proposed cost-effective method achieves complete annotations for constructing an open domain event detection evaluation benchmark. We will analyze how this method reduces the cost of benchmark construction and facilitates the scalability of the benchmark.

D.1 Cost Analysis

Assuming that there are N texts and T event types in the benchmark, then the naive annotation cost is

$$C_{naive}(N, T) = N \times T \times \bar{c}$$

, where \bar{c} represents the average cost of a single annotation of humans or agents such as LLMs.

In our construction pipelines, LLMs are required to identify possible event types in the texts, with an average of \bar{p} types per text. Then, a single-tower model for text similarity is employed to filter out latent event types, with an average of \bar{t} types, from the benchmark that exhibit high similarity to possible event types. Therefore, the cost-effective annotation cost is

$$C(N, T) = N \times \bar{t} \times \bar{c} + N \times \bar{c} + (N \times \bar{p} + T) \times \bar{c}'$$

, where \bar{c}' represents the average cost of computing a text embedding with the single-tower model, and $\bar{t} \approx 10 \ll T$. The overall cost arises from three parts: (1) performing supplementary annotation with \bar{t} latent event types, (2) detecting possible

Test Dataset / Benchmark	Domain	Documents	Tokens	Event Types	Events
ACE05 (Doddington et al., 2004)	News	2519	33824	30	563
MUC-4 (muc, 1992)	News	464	120751	1	464
RAMS (Ebner et al., 2020)	News	910	120162	135	910
PHEE (Sun et al., 2022)	Pharmacovigilance	965	21377	2	1005
Genia2013 (Kim et al., 2013)	Biomedical	139	31341	11	974
MLEE (Pyysalo et al., 2012)	Biomedical	42	10440	21	867
CASIE (Satyapanich et al., 2020)	Cybersecurity	218	63103	5	1268
FewEvent (Deng et al., 2020)	General	2541	80698	99	2541
MAVEN (Wang et al., 2020)	General	5925	150744	168	14190
M2E2 (Li et al., 2020)	Multimedia	901	25467	8	174
MEE-en (Pouran Ben Veyseh et al., 2022)	Wikipedia	1300	106320	16	1745
WikiEvents (Li et al., 2021)	Wikipedia	62	21010	38	379
SEOE (Ours, with $p = 0.5$)	Open Domain	4423	382667	564	28653

Table 5: Basic statistics about sampled datasets used in SEOE. Since our proposed SEOE is an evaluation dataset, we only select the test sets of other datasets for comparison.

Metric	$p = 0.3$	$p = 0.5$	$p = 0.7$
Documents	4423	—	—
Sentences	15095	—	—
Tokens	382667	—	—
Event Types	564	—	—
Avg. Def. Length	109.03	—	—
Events	22333	28653	35015
Avg. Trigger Length	1.27	1.42	1.53
Avg. Events per Doc.	5.04	6.47	7.91
Type’s Avg. Frequency	39.60	50.80	62.08

Table 6: Statistics of SEOE’s benchmark. The benchmark’s supplemental annotations are derived from 10 independent repetitions of annotations along with a nucleus sampling with different threshold p .

event types for each text, and (3) computing embeddings for T integrated event types and \bar{p} possible of each text using the single-tower model. As the single-tower model is typically derived from an off-the-shelf small language model, the third part of the cost is negligible. Therefore, the overall cost can be estimated as follows:

$$C(N, T) \approx N \times \bar{t} \times \bar{c} \ll C_{naive}(N, T)$$

The above analysis demonstrates that our proposed cost-effective method can reduce annotation costs to nearly linear scaling with the number of texts when applied to benchmarks with extensive text data and numerous types.

D.2 Scalability Analysis

Assuming that there are N' new texts and T' new types as the supplement of the benchmark, there are three components of additional cost. The first cost comes from generating fine-grained definitions for new event types, which is linearly proportional to

the number of new types:

$$\Delta C_1 = T' \times \bar{c}_t$$

, where \bar{c}_t represents the average cost of generate the fine-grained definition for a new event type.

The second cost involves annotating new texts using the newly integrated types, which is similar to Section D.1:

$$\begin{aligned} \Delta C_2 &= N' \times (\bar{t} + \bar{t}' + 1) \times \bar{c} \\ &\quad + (N' \times \bar{p} + T') \times \bar{c}' \\ &\approx N' \times (\bar{t} + \bar{t}') \times \bar{c} \end{aligned}$$

, where $\bar{t}' \ll T'$ denotes new latent type from newly added types. This part of the cost is approximately linearly proportional to the number of new texts.

The third cost arises from supplementing the annotated T texts with new types:

$$\Delta C_3 = N \times \bar{t}' \times \bar{c}$$

Let $N_{all} = N + N'$. Then, the total extra cost is:

$$\begin{aligned} \Delta C(N', T') &= \Delta C_1 + \Delta C_2 + \Delta C_3 \\ &\approx T' \times \bar{c}_t + N_{all} \times \bar{t}' \times \bar{c} \\ &\quad + N' \times \bar{t} \times \bar{c} \end{aligned}$$

The above analysis demonstrates that when new texts and types are added to the benchmark, the additional cost is primarily linearly proportional to the number of new types, new texts and annotated texts that need additional annotation. In conclusion, our proposed cost-effective method supports the open benchmark easily scalable.

Test Dataset / Benchmark	Event Type	Top-10%	Top-25%	Bottom-50%	Bottom-25%	Bottom-10%
SEOE (Ours, with $p = 0.3$)	564	37.39	59.54	19.65	7.26	2.22
SEOE (Ours, with $p = 0.5$)	564	36.10	59.36	19.08	6.63	1.89
SEOE (Ours, with $p = 0.7$)	564	35.94	59.89	18.65	6.30	1.81
MAVEN	168	40.09	67.13	11.21	2.67	0.40
RAMS	135	34.28	55.49	20.21	5.71	1.42
FewEvent	99	21.25	56.67	10.74	4.64	1.18
WikiEvents	38	37.46	69.39	10.29	2.37	0.79
ACE05	30	58.43	79.21	6.39	1.59	0.53
MLEE	21	32.75	68.05	7.61	1.96	0.46

Table 7: Cumulative distribution (%) of event types at different frequency levels. Top- $x\%$ and Bottom- $x\%$ refer to the total proportion of event types ranked in the top $x\%$ and bottom $x\%$ based on frequency, respectively. Datasets only with the number of event types over 20 are selected for comparison. Since our proposed SEOE is an evaluation dataset, we only select the test sets of other datasets for comparison.

Annotation Strategy	C1	C2	C3	C4
1 round	52.37	2.26	1.40	24.41
3 rounds w/. $p = 0.5$	50.37	4.60	1.38	24.77
5 rounds w/. $p = 0.5$	49.70	4.37	1.71	27.85
10 rounds w/. $p = 0.3$	54.60	3.56	1.84	26.90
10 rounds w/. $p = 0.5$	49.75	5.03	1.24	29.37
10 rounds w/. $p = 0.7$	44.13	5.62	1.15	30.91

Table 8: The proportion (%) of annotations classified into each correct subtype under different annotation strategies, relative to the total number of annotations.

Annotation Strategy	IC1	IC2	IC3
1 round	2.80	14.95	1.83
3 rounds w/. $p = 0.5$	2.85	14.27	1.75
5 rounds w/. $p = 0.5$	3.17	11.74	1.46
10 rounds w/. $p = 0.3$	1.61	9.43	2.07
10 rounds w/. $p = 0.5$	2.39	10.56	1.65
10 rounds w/. $p = 0.7$	3.70	11.43	3.07

Table 9: The proportion (%) of annotations classified into each incorrect subtype under different annotation strategies, relative to the total number of annotations.

E Prompts

E.1 Benchmark Construction Prompts

In Section 4, we leverage GPT-4o to construct the evaluation benchmark. The prompt used for generating fine-grained definitions for each event type is illustrated with an example in Table 10. The prompt used for identifying possible event types for each data instance is illustrated with an example in Table 11. After recognizing latent event types, the prompt used for generating supplementary annotations is illustrated with an example in Table 12. After nucleus sampling from multiple repeated annotations, the prompt used for merging semantically redundant triggers is illustrated with an example in Table 13.

E.2 Inference and Evaluation Prompts

In our proposed SEOE evaluation framework, we leverage GPT-4o as the automatic evaluation agent to evaluate ODED abilities of open-source and closed-source LLMs. The inference prompt used for LLMs’ inference is illustrated with an example in Table 14. The evaluation prompt, which includes group information in Section 5.1, is illustrated with an example in Table 15.

Prompt Used for Generating Fine-grained Definitions for Each Event Type	
Background	In the context of open domain event extraction, we aim to enhance our understanding of various event types by generating more detailed and fine-grained definitions. This will aid in better classification and recognition of events in text.
Task Description	You are a knowledgeable event extraction specialist tasked with generating a more detailed definition for a specified event type based on its name, ontology, and provided examples of this event type. Your definition should capture the nuances and variations of the event type to facilitate its application in natural language processing tasks. Input: You will receive an event type along with its ontology. Some examples of this event type are provided.
Output Format	Please provide the output in the following Python dictionary format: {"event_type": <str: given_event_type_name>, "detailed_definition": <str: your_generated_definition>}
Data	<p>Event Type Name: Conflict:Attack</p> <p>Ontology: The argument roles of this event type are ['Attacker', 'Target', 'Instrument', 'Place', 'Agent'].</p> <p>Example 0:</p> <p>Text: This isn't a football game, and it's not two drunken cousins [t] fighting [/t] at a family reunion.</p> <p>Trigger: fighting # event type: Conflict:Attack</p> <p>Argument 0: cousins # role: Attacker</p> <p>Argument 1: cousins # role: Target</p> <p>Example 1:</p> <p>Text: When we came across the bridge, the enemy [t] launched [/t] some 155 artillery rounds at us.</p> <p>Trigger: launched # event type: Conflict:Attack</p> <p>Argument 0: enemy # role: Attacker</p> <p>Argument 1: bridge # role: Place</p> <p>Argument 2: us # role: Target</p> <p>Argument 3: rounds # role: Instrument</p> <p>Example 2:</p> <p>Text: luster skipped bail during his trial in January for [t] raping and drugging [/t] three women.</p> <p>Trigger: raping and drugging # event type: Conflict:Attack</p> <p>Argument 0: women # role: Victim</p> <p>Argument 1: luster # role: Agent</p>

Table 10: Prompt used for generating fine-grained definitions for each event type in the integrated ontology. An example of data is illustrated.

Prompt Used for Identifying Possible Event Types	
Background	In the context of open-domain event extraction, we aim to enhance the comprehensiveness of event annotations by supplementing existing annotations with additional potential event types. You are provided with a piece of text and its associated gold annotations, and the goal is to identify potential event types that may not have been annotated but could still exist within the text. This will help ensure that all relevant event types are captured in future analyses.
Task Description	You are tasked with analyzing the provided text and its gold annotations. Your goal is to identify any additional event types that might be present in the text but are not explicitly annotated. For each event type you identify, provide a fine-grained definition that captures the specific nuances of that event type. Before identifying event types, analyze the text to understand it.
Output Format	Provide your answer in the following Python dictionary format: {"text_analysis": "<your_analysis>", "possible_event_types": [{"event_type": "<event_type_name>", "detailed_definition": "<fine_grained_definition>"}, ...]}
Data	<p>Text: How is the world reacting to the war in Iraq?</p> <p>Gold trigger 0: war # event type: Conflict:Attack</p> <p>Detailed definitions of gold event types:</p> <p>Conflict:Attack: The "Conflict:Attack" event type involves an aggressive action or series of actions perpetrated by an Attacker with the intent to cause harm or exert control over a Target. The event may occur at a specific Place and can involve an Instrument, which is any tool or method used to carry out the attack. Additionally, an Agent may be present, indicating an entity responsible for orchestrating or executing the attack. Variations of this event type can range from physical confrontations, such as fights or battles, to assaults involving weapons or other means of inflicting harm, including psychological or chemical methods. The event is characterized by its hostile nature and the presence of a conflict between the involved entities.</p>

Table 11: Prompt used for identifying possible event types for each data instance in the benchmark. An example of data is illustrated.

Prompt Used for Generating Supplementary Annotations	
Background	In the context of open-domain event extraction, we are supplementing existing annotations in text to ensure that all relevant event types are fully captured. You are provided with a piece of text, its associated gold annotations, and a list of possible event types with detailed definitions. Your task is to supplement the existing annotations by identifying additional event types that may be present in the text but not originally annotated.
Task Description	You are tasked with analyzing the provided text and its existing gold annotations. Based on the list of possible event types and their detailed descriptions, you need to supplement the annotations by identifying which of these additional event types are present in the text. Add these event types where appropriate, considering their triggers in the text.
Output Format	Please provide your answer in the following Python dictionary format: {"text_analysis": "<your_analysis>", "supplemented_annotations": [{"event_type": "<event_type_name>", "trigger_word": "<word_or_phrase_in_text>"}, ...]}
Data	<p>Text: How is the world reacting to the war in Iraq?</p> <p>Gold trigger 0: war # event type: Conflict:Attack</p> <p>Detailed definitions of gold event types:</p> <p>Conflict:Attack: The "Conflict:Attack" event type involves an aggressive action ... <i>(Omitted)</i></p> <p>Possible event types with their detailed definitions:</p> <p>Conflict:Demonstrate_2: The event type refers to organized public gatherings ... <i>(Omitted)</i></p> <p>contact.commandorder.correspondence_1: The event type involves the act ... <i>(Omitted)</i></p> <p>Reporting_1: The Reporting event type encompasses instances where ... <i>(Omitted)</i></p> <p>Earnings_and_losses_1: The event type encompasses a wide range of scenarios ... <i>(Omitted)</i></p> <p>Contact.Contact.Broadcast_1: The Contact.Contact.Broadcast event type involves ... <i>(Omitted)</i></p> <p>... <i>(Omitted other 16 event types along with their fine-grained definitions)</i></p>

Table 12: Prompt used for generating supplementary annotations for every piece of data in the benchmark. An example of data is illustrated.

Prompt Used for Merging Semantically Redundant Triggers after Nucleus Sampling	
Background	In the open-domain event extraction task, you are tasked with assisting in the process of merging duplicate event triggers within a dataset. The dataset has undergone multiple rounds of supplementation, resulting in 10 versions of triggers per data point. To ensure robustness and reliability, you have already performed a resampling process to select representative data points. However, during the event extraction task, subtle differences may occur in the mentions and event type definitions across repeated annotations. These differences need to be merged, reducing redundant information. Your job is to assist in merging similar event triggers that refer to the same event, based on their mention and event type definition. The triggers are divided into groups, and each group may contain multiple triggers that describe similar but not identical events. Your goal is to analyze and select the most appropriate trigger for a certain event, ensuring that only one trigger per event is retained.
Task Description	Given the original text and the resampled set of triggers with their corresponding fine-grained event type definitions, you are to perform the following steps: (1) Analyze the original text to understand the overall context and what events it discusses. (2) Analyze each group of triggers. Each group contains multiple triggers that potentially describe similar or related events, but they may not all refer to exactly the same thing. (3) For each group, assess the triggers based on their mentions and event type definitions. You need to determine which trigger best represents the event described and merge the similar triggers by selecting the most appropriate one. Output the merged triggers, ensuring that only the most relevant trigger for each event is retained.
Output Format	{"Text Analysis": "<str: your analysis of the text>", "Group Analysis": [{"str: your analysis of the first group>", "<str: your analysis of the second group>", ...}], "Merged Triggers": [{"Trigger Span": "<str: the selected span in the text>", "Event Type": "<str: the event type corresponding to this trigger>"}, ...]}
Data	<p>Text: The other option, and what the US seems to be trying, is to somehow fudge the election so Sistani does not win, that raise doubts about the outcome, claiming fraud, do anything so that the election causes to results.</p> <p>Group 0:</p> <p>Trigger 0: claiming fraud # event_type: government.vote.violationspreventvote_1</p> <p>Trigger 1: claiming fraud # event_type: contact.prevarication.broadcast_1</p> <p>Trigger 2: fraud # event_type: government.vote.violationspreventvote_1</p> <p>Fine-grained Event Type Definitions: ... <i>(Omitted fine-grained definitions of the above 2 event types)</i></p>

Table 13: Prompt used for merging semantically redundant triggers for every piece of data after nucleus sampling. An example of data is illustrated.

Prompt Used for LLMs' Inference	
Background	You are an event extractor designed to identify the presence of specific events in a given piece of text and to locate the corresponding event triggers in an open-domain setting. Open domain means there are no predefined constraints on the extraction process, and the text may originate from diverse sources and domains. This task requires comprehensive analysis and understanding of the text, as well as the ability to identify and classify events flexibly without reliance on a fixed ontology.
Task Description	In an open-domain setting, your task is to identify all possible triggers and their corresponding event types in the given text. A trigger is the key word or phrase in the text that most explicitly conveys the occurrence of an event. The process includes the following steps: (1) Analyze the text to understand its content and context. (2) Recognize potential event types that may be implied in the text. For each event type, generate a summarized name and provide a detailed description based on your cross-domain knowledge. (3) Extract triggers by identifying their spans in the text and linking them to their respective event types. Note that the text may contain zero, one, or multiple event triggers.
Output Format	Please provide your response in the following Python dictionary format: {"Text Analysis": "<str: your analysis of the text>", "Event Type Generation": {"<event_type_name_1>": "<str: detailed_description>", "<event_type_name_2>": "<str: detailed_description>", ...}, "Extraction Results": [{"Trigger Span": "<str: a span in the text>", "Event Type": "<str: a specific event type from your generation>"}], ...}}
Data	Text: How is the world reacting to the war in Iraq?

Table 14: Prompt used for large language models to do inference. An example of data is illustrated.

Prompt Used for Evaluating Open Domain Event Detection Models	
Background	You are an evaluation master designed to evaluate an open event detection model. In the open-domain event detection task, you aim to evaluate model performance at a semantic level. The evaluation target in open-domain event detection focuses on finding matching pairs of event triggers in the gold annotation set and the prediction set. A trigger refers to the key word or phrase that signals the occurrence of a specific event. Given an original text, a set of golden annotated triggers with event types, and a set of predicted triggers with event types, your goal is to match triggers with similar event types from both sets based on semantic equivalence. This evaluation does not require exact matches in trigger mention or event type definitions; rather, it focuses on identifying pairs of triggers that refer to the same event in the text and have semantically similar event type definitions.
Task Description	You are provided with an original text, a set of golden annotated triggers (with corresponding event types and definitions), a set of predicted triggers (also with event types and definitions), and a set of similar event type definitions used to assist in distinguishing event definitions. Your task is to: (1) Analyze the main event or events discussed in the original text. (2) Try to understand each golden trigger and each predicted trigger, along with their corresponding event type definitions. (3) Find several possible matching pairs of triggers that refer to the same event and have event type definitions that refer to the same event type. Then, analyze the event type definitions of two triggers' event types for each pair, and judge whether each pair can be regarded as a matching pair or not. (4) Provide the matching pairs of triggers' index as your final output.
Output Format	Please provide your response in the following Python dictionary format: {"text_analysis": "<str: your analysis>", "golden_triggers_understanding": [{"<str: understanding for golden trigger 0>", "<str: understanding for golden trigger 1>", ...}], "predicted_triggers_understanding": [{"<str: understanding for predicted trigger 0>", "<str: understanding for predicted trigger 1>", ...}], "possible_matching_pairs": [{"golden_trigger": "<str: golden trigger mention # its event type>", "predicted_trigger": "<str: predicted trigger mention # its event type>", "analysis_and_judgment": "<str: your analysis and judgment>"}], ... }, "final_output": [{"Gold_trigger_index": <int>, "Pred_trigger_index": <int>}, ...]}
Data	Text: How is the world reacting to the war in Iraq? Gold_trigger_0: war # event_type: Conflict:Attack_1 Gold_trigger_1: reacting # event_type: Response_1 Pred_trigger_0: reacting # event_type: Global Response Detailed definition of the above event types: ... <i>(Omitted fine-grained definitions of the above event types)</i> Similar event of the above event types: ... <i>(Omitted fine-grained definitions of event types within the same group as the above golden event types.)</i>

Table 15: Prompt used for evaluating open domain event detection models. An example of data is illustrated.