Beyond Exact Match: Semantically Reassessing Event Extraction by Large Language Models

Anonymous ACL submission

Abstract

Event extraction has gained extensive research attention due to its broad range of applications. However, the current mainstream evaluation method for event extraction relies on token-level exact match, which misjudges numerous semantic-level correct cases. This reliance leads to a significant discrepancy between the evaluated performance of models under exact match criteria and their real performance. To address this problem, we propose a reliable automatic evaluation framework that accurately assesses event extraction results at semantic-level instead of token-level, named RAEE. Specifically, RAEE leverages large language models (LLMs) as evaluation agents, incorporating an adaptive mechanism to achieve adaptive evaluations for precision and recall of triggers and arguments. Extensive experiments demonstrate that: (1) RAEE achieves a very high correlation with human judgments; (2) after reassessing 14 models, including advanced LLMs, on 10 datasets, there is a significant performance gap between exact match and RAEE. The exact match evaluation significantly underestimates the performance of existing event extraction models, and in particular underestimates the capabilities of LLMs; (3) fine-grained analysis under RAEE evaluation reveals insightful phenomena worth further exploration. The evaluation toolkit of our proposed RAEE will be publicly released.

1 Introduction

Event extraction (EE) aims to detect event instances as well as all of their participants and attributes by analyzing and identifying mentions of semantically defined entities and relationships within them (Grishman, 2015; Wei and Wang, 2019). It is widely divided into two subtasks, **E**vent **D**etection (ED) and **E**vent **A**rgument **E**xtraction (EAE). While ED aims to identify all triggers that most explicitly convey the occurrence of an event, EAE extracts arguments with semantic roles corresponding to the trigger. Since EE benefits a broad range of applications (Zhang et al., 2020; Han et al., 2021), it has gained extensive research attention in recent years (Lin et al., 2020; Lu et al., 2022; Yan et al., 2023). Additionally, as large language models (LLMs) are increasingly used as agents (Chen et al., 2024; Liu et al., 2023a), the requirement for accurate event extraction becomes more important, which helps in effectively tracking and labeling the actions of agents in real-world scenarios (Park et al., 2023).

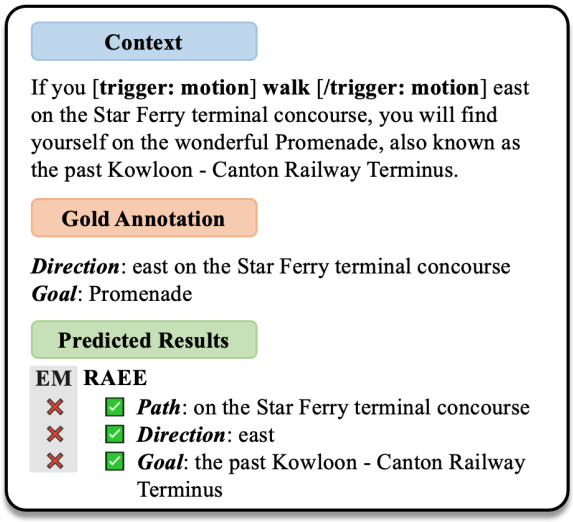

Figure 1: One case of EAE is misjudged by exact match (EM) evaluation but reassessed as correct by our proposed RAEE evaluation. With a given trigger (***walk***), EAE aims to extract corresponding arguments with semantic roles.

Currently, the most widely-used evaluation method for EE is token-level exact match (EM)[1], which assesses whether the predicted triggers and arguments are identical to the annotated ground-truth in benchmarks (Hsu et al., 2022; Lu et al., 2022; Hsu et al., 2023a). However, this EM evaluation often misjudges a significant number of cases

[1]For the sake of brevity, we will refer to exact match evaluation as EM in the following content.

that are semantically correct or convey the same meaning as ground-truth but cannot match it literally (Li et al., 2021). For example, one case of EAE task, which extracts event arguments corresponding to a given trigger, is illustrated in Figure 1. The predicted arguments, *Path: on the Star Ferry terminal concourse* and *Direction: east*, are not the same as ground-truth and are therefore evaluated as failed extractions under EM. However, considering the context, both predictions can be regarded as semantically correct extractions. Also, co-reference instances are misjudged by EM, as shown in the example where *Promenade* and *the past Kowloon - Canton Railway Terminus* refer to the same place. Consequently, models that may appear to perform poorly under EM actually have strong performance, indicating that the capabilities of existing EE models might be significantly underestimated. This issue is particularly pronounced in generative models, especially in LLMs (Huang et al., 2024), where the generations often cannot exactly match ground-truth.

To address this severe issue, we propose a reliable automatic evaluation framework designed to **R**e**A**ssess **E**vent **E**xtraction models at semantic-level instead of token-level, named **RAEE**. Specifically, advanced LLMs are utilized to assess the quality of extractions by considering their semantic correctness within their context, such as the source document or sentence. Also, an adaptive prompting mechanism is designed to guide LLMs in generating adaptive evaluation results for precision and recall of triggers and arguments, in a chain-of-thought manner (Wei et al., 2023).

Meta-evaluation results of 8 models (including advanced LLMs) on 10 datasets demonstrate that RAEE achieves a strong correlation with human judgments and significantly outperforms EM. It indicates that our proposed RAEE is a more reliable EE evaluation framework. After the reliable meta-evaluation, we conduct a comprehensive reassessment for 14 models on 10 datasets using our proposed RAEE. Extensive experiments reveal a significant performance gap between EM and RAEE, suggesting that existing baselines, particularly LLMs, are greatly underestimated. Although LLMs still slightly underperform smaller fine-tuned models under RAEE, the performance gap narrows considerably. Furthermore, the ranking order among these baselines shifts significantly after reassessing under RAEE on most datasets. For example, on ED task, Tagprime-C (Hsu et al., 2023a) and DEGREE (Hsu et al., 2022) significantly outperform other baselines that perform better under EM. Finally, a fine-grained analysis of correct and incorrect cases evaluated by RAEE uncovers insightful phenomena worth future exploration. For example, while most spans are accurately matched, their event and role types are still incorrect under RAEE evaluation. In summary, the contributions of our study are as follows:

- **RAEE**: We propose a reliable automatic evaluation framework to assess event extraction at semantic-level by leveraging advanced LLMs with an adaptive mechanism, showing strong correlations with human judgments.
- **Comprehensive evaluation**: We use our proposed RAEE to comprehensively reassess the performance of previous works and LLMs for event extraction. Experimental results reveal that the capabilities of event extraction models are significantly underestimated, especially of generative models and LLMs.
- **Fine-grained analysis**: We conduct a fine-grained analysis to uncover the correct and incorrect cases evaluated by RAEE, highlighting areas worth further exploration.

2 Related Works

2.1 Evaluations on Event Extraction

Current evaluations of EE typically use precision, recall, and f1-score metrics to evaluate the span and the type of extracted instances, and this approach is consistent across ED and EAE. The majority of research relies on EM evaluation (Lin et al., 2020; Ma et al., 2022; Hsu et al., 2023b), and some studies employ language toolkits to assist with token-level evaluation (Li et al., 2021). Additionally, some works modify the existing evaluation frameworks by incorporating different averaging methods for calculating f1-score (Zheng et al., 2021) or introducing additional metrics from downstream applications (Peng et al., 2023a). The consistency and fairness of variable alignment with model outputs are also emphasized in recent studies (Peng et al., 2023b; Huang et al., 2024). However, these works uniformly adopt EM when determining span matches, and overlook instances where the semantically correct span cannot exactly match ground-truth. To address this oversight, we propose a semantic-level evaluation framework and provide a more reliable evaluation of model performances.

2.2 LLMs as Judge Models

There is an emerging trend in the research community to leverage LLMs for automatically evaluating the quality of generations, thereby reducing the significant cost of human annotations (Lan et al., 2024; Lin et al., 2024; Luo et al., 2023; Fu et al., 2023; Wang et al., 2023). Extensive studies (Fu et al., 2023; Wang et al., 2023) have shown that advanced LLMs, such as GPT-4, achieve very close correlations with human judgments across various generative NLP tasks, including dialogue generation (Fu et al., 2023), text summarization (Liu et al., 2023b) and structured generation (Wang et al., 2023).

In this paper, we utilize advanced LLMs to emulate human annotators in automatically evaluating the quality of EE. Advanced LLMs achieve performance that is closely aligned with human judgments while incurring significantly lower costs (Lan et al., 2020), making our proposed RAEE a scalable and reliable automatic evaluation method for EE.

3 Automatic Reassessing Framework

To address the misjudgments caused by EM, we propose an automatic evaluation framework, RAEE, to reassess existing event extraction methods more accurately. In this section, we will introduce: (1) how RAEE differs from EM in Section 3.1; (2) how RAEE automatically achieves interpretable and adaptive evaluations in Section 3.2; (3) the reliability of RAEE in Section 3.3.

3.1 Semantic-level F1-score

Figure 2 shows the difference between RAEE and EM. Given the context C, predicted results from an EE model $\mathbf{P} = \text{Model}(C)$, and gold annotations from human $\mathbf{G} = \text{Human}(C)$. Evaluation frameworks are required to identify the correctness of each prediction and the recall of each annotation.

In EM, the calculation of true positive cases (TP) is performed by taking the intersection of $\mathbf{P}$ and $\mathbf{G}$. Then, precision and recall are computed as follows:

$$\mathbf{TP} = |\mathbf{P} \cap \mathbf{G}|, \quad \mathbf{p}_{\text{EM}} = \frac{\mathbf{TP}}{|\mathbf{P}|}, \quad \mathbf{r}_{\text{EM}} = \frac{\mathbf{TP}}{|\mathbf{G}|}.$$

Different from EM, RAEE leverages a LLM agent to take the context into account to evaluate predictions and annotations, and get a correct prediction set $\mathbf{P_c}$ and a recalled annotation set $\mathbf{G_r}$:

$$\mathbf{P_c} = \mathbf{LLM}(\mathbf{P}|\mathbf{C}, \mathbf{G}), \quad \mathbf{G_r} = \mathbf{LLM}(\mathbf{G}|\mathbf{C}, \mathbf{P}).$$

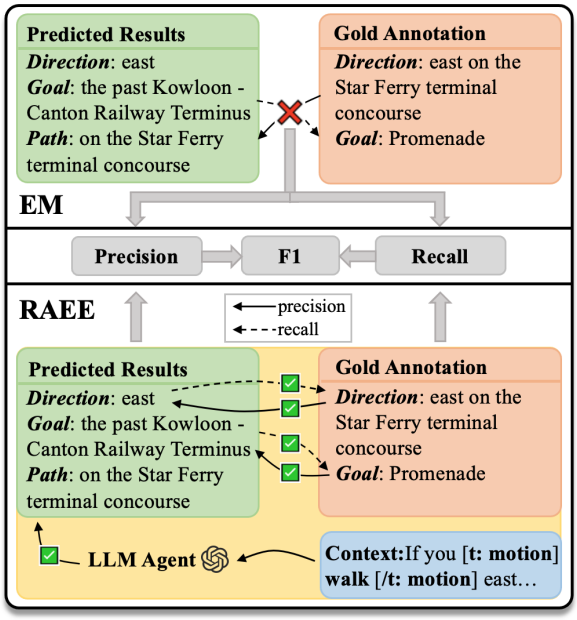

Figure 2: Comparison of EM and RAEE evaluation using EAE as an example (consistent with ED). RAEE utilizes a LLM to assess event extraction results from both precision and recall perspectives. The source of the arrows indicates the basis or reason for judgment, with solid arrows representing the precision evaluation process and dashed arrows representing the recall evaluation process.

Then, semantic-level f1-score can be computed without calculating TP:

$$\mathbf{p} = \frac{|\mathbf{P_c}|}{|\mathbf{P}|}, \quad \mathbf{r} = \frac{|\mathbf{G_r}|}{|\mathbf{G}|}, \quad \mathbf{f1}_{\text{RAEE}} = \frac{\mathbf{2 * p * r}}{\mathbf{p + r}}.$$

Since human evaluation is the most reliable golden standard (Wang et al., 2023; Fu et al., 2023), the most accurate results for $\mathbf{P_c}$ and $\mathbf{G_r}$ should ideally be obtained through human annotation. However, due to the high cost, lengthy duration, and lack of reproducibility associated with human evaluation (Lan et al., 2024, 2020), the advanced LLM, GPT-4o, is utilized as a reliable proxy for human evaluation in RAEE. Specifically, LLMs are equipped with an adaptive mechanism to evaluate the quality of event extraction results by considering the semantic correctness within context in a chain-of-thought manner.

3.2 Adaptive Evaluation by LLMs

Human annotators might have slightly different subjective perspectives when evaluating, which could lead to differences in assessment (Lan et al., 2024; Kim et al., 2023). Therefore, for both ED and EAE tasks, we manually identify several key

criteria of evaluation bias, such as evaluating pronouns in co-reference as correct extractions, and prompt LLMs to follow them to judge the quality of event extraction results step by step. More details about our human-annotated key criteria of evaluation bias and reassessment prompts can be found in Appendix C and Appendix D.1, respectively.

It should be noted that these key criteria are optional and adaptive, contributing to the adaptive mechanism in prompts. This allows researchers to tailor the evaluation criteria of RAEE to their specific needs in different scenarios. For example, some datasets provide separate annotations for co-references to assist in the annotation of event extraction (Li et al., 2021) but some do not (Ebner et al., 2020; Satyapanich et al., 2020). To achieve fair comparison in experiments, the evaluation criteria remain the same for all the evaluated models in the following reassessment experiments.

Task	-	Agree. ↑	Spearman ↑
ED-p	Human-Average	89.08	74.18
	GPT4o-Human	76.89	49.09
	- w/o. Adaptive	78.99	47.25
	EM-Human	26.05	——
ED-r	Human-Average	96.87	92.64
	GPT4o-Human	87.17	74.72
	- w/o. Adaptive	84.19	69.75
	EM-Human	71.05	——
EAE-p	Human-Average	93.61	85.47
	GPT4o-Human	88.13	72.55
	- w/o. Adaptive	77.28	53.16
	EM-Human	63.49	——
EAE-r	Human-Average	96.96	94.00
	GPT4o-Human	86.10	75.27
	- w/o. Adaptive	86.42	76.08
	EM-Human	72.55	——

Table 1: Agreement percent and Spearman correlation with human judgments. -p and -r represent the evaluation of precision and recall, respectively. All Spearman correlation scores are associated with the p-value $<$ 0.05. It denotes that LLM achieves a moderate and strong correlation with human judgments. Spearman correlation scores of EM cannot be computed due to the failure of basic correlation.

3.3 RAEE Correlation with Human Evaluation

To prove the effectiveness of our proposed RAEE evaluation framework and adaptive evaluation, we conduct a meta-evaluation to analyze the correlation between RAEE and human judgments. Specifically, we invite 3 professional event extraction annotators to evaluate 200 sampled pieces of data for both ED and EAE tasks on precision and recall metrics, as 800 evaluation pieces in total. These annotators are given the same prompts as those given to LLM. With reassessing results from 3 humans and a LLM (GPT-4o[2]), we calculate two widely-used inter-annotator agreement (IAA) metrics among humans and LLM: percent agreement (Zheng et al., 2023; Liu et al., 2021) and Spearman correlation (Lan et al., 2020, 2024; Fu et al., 2023). The experimental results are shown in Table 1.

3.3.1 Correlation on Recall Task

While LLM correlation with human judgments shows a noticeable drop (average 18.33% Spearman correlation and 10.28% agreement percent), LLM still achieves a strong correlation[3] with human judgments on Agreement percent and Spearman correlation, *i.e.,* over 86% agreement percent, which significantly outperforms EM with over 14.84% agreement percent. It also suggests that while LLM emulates human judgments well and can be regarded as an automatic evaluation agent, there is still room for further refinement.

3.3.2 Correlation on Precision Task

LLMs achieve a strong correlation with human judgments, over 76% agreement percent and 49% Spearman correlation of ED-p (Lan et al., 2020; Fu et al., 2023). Furthermore, the average correlation scores between LLMs and human judgments for ED-p are significantly lower than those in the other tasks, indicating that semantic evaluation in this task is more challenging. Our investigation reveals two main reasons for this phenomenon: (1) Different datasets within the same domain have personalized definitions for event types with similar names; (2) As for datasets across different domains, there are discrepancies in how events are defined as being triggered. When sampling data from multiple datasets, these subtle definitional differences can lead to confusion among evaluators when determining whether trigger predictions are accurate, leading to lower consistency in ED-p. In contrast to ED-p, evaluating recall is simpler since it only requires determining whether gold annotations are retrieved, without judging the accuracy of event types. Also, evaluating event arguments is more straightforward, as role types of event arguments

[2]According to https://openai.com/index/hello-gpt-4o/, studies show that gpt-4o performs best in text evaluation.

[3]The following metrics value can be considered as strong correlation in the validation of past works(Lan et al., 2020; Fu et al., 2023).

are clearly defined across datasets. The above analysis explains the relatively higher IAA observed in the three tasks other than ED-p.

3.3.3 Ablation of Adaptive Mechanism

Moreover, it is observed that the adaptive mechanism improves Spearman correlation between RAEE and human average in most cases, *i.e.,* over 6.34% Spearman correlation absolute improvements on average. Such improvements are a result of the pre-defined criteria helping to reduce the randomness in the LLM evaluation, which cannot be detected by agreement percent evaluation.

Summary While LLM slightly underperforms human judgments, it significantly outperforms EM and achieves a strong correlation with human judgments. It is important to note that RAEE is a **trade-off** solution by considering the evaluation cost and reliability, *i.e.,* RAEE can reliably assess the performance of event extraction results with affordable cost. As shown in Table 2, the average time and financial cost of GPT-4o on an evaluated model and a dataset are much smaller than those of pure human annotation, proving the scalability and reliability of LLM judging models in our proposed RAEE framework.

Evaluation Cost	Time (hours)	Financial ($)
RAEE + GPT-4o	≈ 0.23	≤ 5
RAEE + Human	≈ 19.22	≈ 182.32

Table 2: The average evaluation cost of human annotation and GPT-4o for an evaluated model on a dataset.

4 Re-Assessment Experiments

The reliability of RAEE with GPT-4o as the judging agent has been proven. In this section, we reassess the performance of existing EE models in numerous datasets across various domains by using our proposed RAEE evaluation framework.

4.1 Experimental Settings

Tasks and Metrics ED and EAE tasks are evaluated respectively. Following (Lin et al., 2020), we focus on trigger classification (TC) and argument classification (AC). The semantic-level f1-score for TC and AC is described in Section 3.1, using specialized prompts to assess both precision and recall.

Reassessed Models We reassess the overall 8 state-of-the-art event extraction models with different frameworks (5 models for ED and EAE tasks). These models can be broadly classified into two categories: (1) Discriminative models that classify tokens of the text into predefined categories, including Tagprime-C (Hsu et al., 2023a), OneIE (Lin et al., 2020), Query&Extract (Wang et al., 2022), CEDAR (Li et al., 2023a); (2) Generative models that produce event-related outputs directly, including BartGen (Li et al., 2021), PAIE (Ma et al., 2022), DEGREE (Hsu et al., 2022), AMPERE (Hsu et al., 2023b). All models train and inference under the unified TextEE framework (Huang et al., 2024), with implementation details shown in Appendix E.

Event Datasets We test the performance of event models using a total of 10 datasets covering diverse domains, including news of ACE05-en (Doddington et al., 2004) and RAMS (Ebner et al., 2020), general domain of FewEvent (Deng et al., 2020) and GENEVA (Parekh et al., 2023), Wikipedia of MEE-en (Pouran Ben Veyseh et al., 2022) and WikiEvents (Li et al., 2021), biomedical domain of Genia2011 (Kim et al., 2011), multimedia of M^2E^2 (Li et al., 2020), cybersecurity of CASIE (Satyapanich et al., 2020), and pharmacovigilance of PHEE (Sun et al., 2022). The statistics of these datasets are provided in Appendix B.

4.2 Main Results

Table 3 and 4 show both EM evaluations and RAEE reassessment results for ED and EAE tasks, respectively. Notably, across all models, RAEE reassessment shows a significant improvement in metrics compared to EM evaluations. On average, the precision of all models increased by more than 20%, with recall improvement exceeding 10% for ED and 15% for EAE. It indicates that EM evaluation severely underestimates model performance by overlooking a substantial number of semantically correct predictions, while RAEE provides a more accurate evaluation of model performance across diverse datasets.

Moreover, as shown in Table 5, the performance gap between RAEE and EM is significantly larger for generative models compared to discriminative models, indicating that EM introduces a greater bias for generative models.

Furthermore, the rank of models across different datasets changes considerably under RAEE. For instance, in the ED task, DEGREE ranks lower

Model	ACE05			Genia2011			CASIE			PHEE		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1
*Token-level Evaluation under **Exact Match***												
Tagprime-C	62.69	74.60	68.13	75.16	71.91	73.50	68.29	71.85	70.02	72.56	70.39	71.46
OneIE	70.94	70.69	70.82	75.52	72.30	73.87	69.68	72.32	70.98	74.86	67.10	70.77
Query&Extract	66.44	68.92	67.65	68.10	69.20	68.65	47.90	58.36	52.61	46.90	67.10	55.21
CEDAR	60.30	71.76	65.53	72.19	64.88	68.34	70.05	68.06	69.04	69.88	71.49	70.68
DEGREE	60.41	73.18	66.18	54.42	70.97	61.60	57.43	65.85	61.35	68.36	68.49	68.43
*Semantic-level Evaluation under **RAEE***												
Tagprime-C	85.82	81.71	83.71	94.15	88.39	91.18	92.65	93.53	93.09	87.77	78.51	82.88
OneIE	89.30	77.98	83.26	93.34	86.82	89.97	93.77	94.09	93.93	89.10	74.63	81.22
Query&Extract	87.16	76.38	81.41	91.88	85.90	88.78	81.75	91.88	86.52	66.83	75.52	70.91
CEDAR	79.10	78.15	78.63	92.05	82.22	86.86	93.83	90.46	92.11	84.21	78.81	81.42
DEGREE	82.70	81.53	82.11	88.95	87.84	88.39	92.71	90.54	91.61	86.07	80.20	83.03

Model	Fewevent			MEE-en			M2E2			Average		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1
*Token-level Evaluation under **Exact Match***												
Tagprime-C	63.62	66.00	64.79	79.27	82.18	80.70	46.35	62.07	53.07	66.85	71.28	68.81
OneIE	68.38	62.65	65.39	78.22	82.52	80.31	52.10	50.00	51.03	69.96	68.23	69.02
Query&Extract	67.42	60.68	63.88	77.22	82.18	79.62	47.70	65.52	55.21	60.24	67.42	63.26
CEDAR	44.18	62.26	51.68	81.71	79.37	80.52	45.54	52.87	48.94	63.41	67.24	64.96
DEGREE	60.33	59.78	60.05	78.06	78.68	78.37	43.46	59.20	50.12	60.35	68.02	63.73
*Semantic-level Evaluation under **RAEE***												
Tagprime-C	88.24	73.87	80.42	93.42	86.99	90.09	94.42	64.94	76.96	90.93	81.13	85.48
OneIE	90.03	69.85	78.67	93.21	86.65	89.81	96.41	53.45	68.77	92.17	77.64	83.66
Query&Extract	89.37	67.57	76.96	92.95	87.11	89.93	93.72	69.54	79.84	86.24	79.13	82.05
CEDAR	61.83	69.30	65.35	94.40	83.27	88.48	89.11	57.47	69.88	84.93	77.10	80.39
DEGREE	86.38	67.69	75.90	93.12	84.81	88.77	96.20	68.39	79.95	89.45	80.14	84.25

Table 3: Evaluation under EM versus reassessment under RAEE on ED task. The values in the Average column are derived from the mean of the values across 7 datasets. For each metric in each evaluation framework, we highlight the highest value using a dark shade, and the second-highest value using a light shade.

than OneIE and CEDAR on average under EM evaluation, while DEGREE outperforms these two models under RAEE. It suggests that EM evaluation distorts the true performance rank of models, leading to misconceptions about the model performance.

4.3 Fine-grained Reason Analysis

We further conduct a fine-grained analysis to uncover the main reasons behind the misjudgments of EM, as well as the primary failure modes of existing event extraction models under our proposed RAEE evaluation framework. Specifically, 200 sampled pieces of data for both ED and EAE tasks are evaluated through multiple human experts.

Reasons of EM Misjudgements As for instances evaluated as incorrect under EM but reassessed as correct under RAEE, we summarize 4 main categories of reasons for ED and EAE tasks, through human annotations: (1) ***With CoreWord:*** Compared to the ground-truth, the predicted results contain the core words and only have missing or redundant words that do not affect the overall semantics; (2) ***Co-reference:*** There is a relationship between the predicted result and a certain ground-truth refer to the same entity or concept; (3) ***Correct:*** While no match is found in ground-truth, these instances are still reassessed as correct; (4) *Other non-representative reasons*.

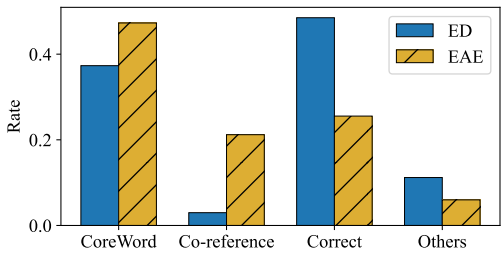

Figure 3: Distribution of the reasons that misjudges by EM evaluation method on ED and EAE tasks.

As shown in Figure 3, it is observed that the most misjudgments for EM evaluation in ED and EAE tasks are *CoreWord* and *Correct*. It indicates that

Model	ACE05			Genia2011			CASIE			PHEE		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1
*Token-level Evaluation under **Exact Match***												
Tagprime-C	75.23	70.05	72.55	83.46	68.79	75.42	66.39	66.07	66.23	71.69	45.84	55.92
BartGen	72.19	67.39	69.71	78.10	64.58	70.70	59.68	60.52	60.10	58.17	43.49	49.77
PAIE	74.14	70.29	72.16	81.07	67.84	73.87	65.13	62.90	64.00	72.99	74.34	73.66
DEGREE	74.30	67.39	70.68	70.08	57.95	63.44	55.29	51.32	53.23	61.42	47.63	53.65
AMPERE	69.87	64.13	66.88	71.45	56.89	63.35	59.08	50.21	54.29	60.94	44.54	51.46
*Semantic-level Evaluation under **RAEE***												
Tagprime-C	92.72	79.11	85.37	92.30	72.35	81.12	92.61	83.15	87.62	89.89	81.57	85.53
BartGen	91.34	79.59	85.06	91.26	73.79	81.60	86.21	80.74	83.39	84.16	79.26	81.64
PAIE	92.35	81.52	86.60	90.56	76.41	82.89	92.01	83.91	87.78	89.22	92.73	90.94
DEGREE	93.35	77.29	84.57	86.87	68.14	76.37	88.16	71.82	79.16	86.75	82.37	84.50
AMPERE	88.96	74.88	81.32	87.71	67.44	76.25	89.31	66.98	76.55	85.89	80.23	82.97

Model	WikiEvents			RAMS			GENEVA			Average		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1
*Token-level Evaluation under **Exact Match***												
Tagprime-C	66.80	63.69	65.21	50.34	45.02	47.53	78.25	78.99	78.62	70.31	62.64	65.93
BartGen	70.24	63.69	66.80	48.70	43.37	45.88	64.20	63.69	63.94	64.47	58.10	60.99
PAIE	70.71	65.63	68.08	51.94	51.06	51.50	70.43	68.41	69.41	69.49	65.78	67.52
DEGREE	57.27	51.26	54.10	45.37	44.60	44.98	64.88	66.17	65.52	61.23	55.19	57.94
AMPERE	60.42	50.10	54.78	47.87	46.63	47.24	65.41	66.62	66.01	62.15	54.16	57.71
*Semantic-level Evaluation under **RAEE***												
Tagprime-C	93.32	76.36	83.99	81.31	70.14	75.31	93.65	92.75	93.20	90.83	79.35	84.59
BartGen	93.99	76.94	84.61	81.49	68.72	74.56	90.34	83.36	86.71	88.40	77.49	82.51
PAIE	92.74	79.84	85.81	81.23	74.89	77.93	93.25	88.77	90.96	90.19	82.58	86.13
DEGREE	88.48	70.74	78.62	80.22	72.87	76.37	91.43	86.14	88.70	87.89	75.62	81.18
AMPERE	93.61	67.64	78.53	81.70	72.87	77.03	91.56	86.64	89.03	88.39	73.81	80.24

Table 4: Evaluation under EM versus reassessment under RAEE on EAE task. The values in the Average column are derived from the mean of the values across 7 datasets. For each metric in each evaluation framework, we highlight the highest value using a dark shade, and the second-highest value using a light shade.

Task	Model Category	Average Δ f1
ED	Discriminative Models	16.38
	Generative Models	**20.52**
EAE	Discriminative Models	18.66
	Generative Models	**21.47**

Table 5: Average f1-score difference between EM and RAEE. It is found that generative models are significantly more underestimated compared to discriminative models.

current event extraction models tend to retain core words while expanding or omitting some less relevant words without changing the semantic meaning of the extracted results. Additionally, it also suggests that annotations of existing datasets may not be comprehensive, as there are still a significant number of extracted results that are reasonable and correct but are not included in the ground-truth.

Failure Modes of Existing Models Under RAEE As for instances evaluated as incorrect under both EM and RAEE, we also summarize 4 kinds of reasons for ED and EAE tasks, through human annotations: (1) ***Without CoreWord:*** Compared to the ground-truth, the predicted results have missing core words that convey the primary semantics; (2) ***WrongType:*** Even though there is a semantic-level match of span in the ground-truth, the predicted results have an incorrect type of event or role; (3) ***Incorrect:*** These instances are reassessed as incorrect and they are not included in the ground-truth; (4) *Other non-representative reasons*.

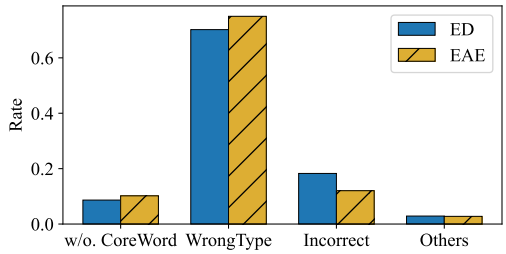

Figure 4: Distribution of failure modes under our proposed RAEE evaluation framework on ED and EAE.

As shown in Figure 4, the most frequent failure

mode of existing models on ED and EAE tasks is *WrongType*. This phenomenon indicates that although existing models could locate the spans correctly, their types are usually misclassified. It suggests that future works should focus on improving the capabilities of models in understanding and classifying the types of triggers and arguments.

Model	RAEE p	RAEE r	RAEE f1	EM f1
Tagprime-C	**93.75**	**86.38**	**89.91**	**78.26**
DEGREE	89.31	85.91	87.58	72.48
GPT-4o	**79.08**	53.05	63.50	32.76
- 2 shots	77.66	48.35	59.60	33.17
- 4 shots	78.77	49.76	60.99	28.23
- 8 shots	74.41	**61.50**	**67.34**	**37.79**
GPT-3.5	55.69	45.07	49.82	13.94
Claude-3-Opus	76.31	40.84	53.21	13.69
Llama-3.1-70B	72.60	54.46	62.23	16.66
Qwen2-70B-Chat	61.53	55.39	58.30	14.47
Mixtral-8x7B	45.84	50.23	47.94	2.72

Table 6: LLMs versus Top-2 state-of-the-art smaller fine-tuned models on ED task under RAEE evaluation.

Model	RAEE p	RAEE r	RAEE f1	EM f1
PAIE	89.90	**86.00**	**87.91**	**67.76**
Tagprime-C	**91.93**	81.81	86.58	65.60
GPT-4o	**83.24**	**73.19**	**77.89**	15.89
- 2 shots	82.49	69.56	75.48	16.36
- 4 shots	81.75	69.02	74.85	16.38
- 8 shots	80.02	69.92	74.63	15.57
GPT-3.5	62.87	64.13	63.49	10.26
Claude-3-Opus	78.82	67.39	72.66	**20.22**
Llama-3.1-70B	60.62	69.56	64.78	12.86
Qwen2-70B-Chat	79.95	68.11	73.56	16.99
Mixtral-8x7B	62.17	65.94	64.00	11.76

Table 7: LLMs versus Top-2 state-of-the-art smaller fine-tuned models on EAE task under RAEE evaluation.

5 Do LLMs Solve EE Task?

Under EM evaluation system, LLMs have a significant performance gap with state-of-the-art small fine-tuned models on EE task (Huang et al., 2024). In this section, we also reassess the capabilities of LLMs. Specifically, 6 advanced open-source and closed-source LLMs are included: GPT-4o, GPT-3.5-turbo, Claude-3-Opus, Llama-3.1-70B-Chat, Qwen2-70B-Chat (Yang et al., 2024) and Mixtral-8X7B-instruct (Jiang et al., 2024). The well-selected prompts for inference on both ED and EAE tasks are shown in Appendix D.2.

Table 6 and 7 demonstrate extraction results of LLMs and Top-2 smaller fine-tuned models under EM and RAEE evaluation on ED and EAE tasks, respectively. It is found that LLMs generally underperform smaller fine-tuned models under EM evaluation, with significantly lower f1-score on both ED and EAE tasks. However, LLMs show remarkable improvements under RAEE evaluation, indicating that the event extraction capabilities of LLMs have been greatly underestimated. Notably, f1-score improvement of LLMs on EAE task, over 51.92%, is higher than that on ED task, over 30.74%. Our human annotations over hundreds of cases reveal that LLMs are more likely to generate out of predefined event type labels in ED than role type labels in EAE, as reflected in their relatively lower recall scores of ED.

Additionally, we also conduct few-shot experiments with the best-performing LLM, GPT-4o. It is found that as the number of few-shot examples increases, the performance of the model fluctuates on ED task and declines on EAE task. Our human annotations reveal that these results are heavily influenced by the quality of selected shots. With data to be extracted and sampled from datasets in a diverse range of domains, it is more likely to mislead the model with a few shots from only a few specific domains. We will leave a more detailed analysis of few-shot prompting strategies in LLMs for our future work.

6 Conclusion and Future Works

In this paper, we propose a reliable automatic evaluation framework, RAEE, for assessing EE at semantic-level rather than token-level of exact match evaluation. Extensive experiments demonstrate that GPT-4o achieves a strong correlation with human judgments within our RAEE framework, establishing RAEE as a reliable automatic evaluation framework for EE. Additionally, we reassess the performance of numerous prior works and advanced LLMs in EE using our proposed RAEE framework. Extensive experiments indicate that the capabilities of prior works and LLMs are significantly underestimated. Finally, the fine-grained reason analysis reveals several intriguing phenomena worth further exploration.

In the future, we will conduct a more comprehensive evaluation for EE to reassess the performance of existing models. Besides, we will extend the evaluation of model performance with RAEE from closed domains to the open domain.

References

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.

Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, and Feng Zhao. 2024. T-eval: Evaluating the tool utilization capability of large language models step by step. *Preprint*, arXiv:2312.14033.

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 *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 151–159, New York, NY, USA. Association for Computing Machinery.

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).

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.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Ralph Grishman. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7543–7559, Online and Punta Cana, Dominican Republic. 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.

I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2023a. TAGPRIME: A unified framework for relational structure extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12917–12932, Toronto, Canada. Association for Computational Linguistics.

I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, and Nanyun Peng. 2023b. AMPERE: AMR-aware prefix for generation-based event argument extraction model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10976–10993, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. Textee: Benchmark, reevaluation, reflections, and future challenges in event extraction. *Preprint*, arXiv:2311.09562.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–37.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-ling Mao. 2024. Criticbench: Evaluating large language models as critic. *arXiv preprint arXiv:2402.13764*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural

language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

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.

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.

Sha Li, Qiusi Zhan, Kathryn Conger, Martha Palmer, Heng Ji, and Jiawei Han. 2023a. GLEN: General-purpose event detection for thousands of types. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2823–2838, Singapore. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

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.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. *arXiv preprint arXiv:2402.14809*.

Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *AAAI Conference on Artificial Intelligence*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023a. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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.

Liangchen Luo, Zi Lin, Yinxiao Liu, Lei Shu, Yun Zhu, Jingbo Shang, and Lei Meng. 2023. Critique ability of large language models. *arXiv preprint arXiv:2310.04815*.

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, Dublin, Ireland. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023a. OmniEvent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517, Singapore. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023b. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9206–9227, Toronto, Canada. Association for Computational Linguistics.

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.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8749–8757.

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.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022. Query and extract: Refining event extraction as type-oriented binary decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xiang Wei and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, PP:1–1.

Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2023. UTC-IE: A unified token-pair classification architecture for information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4096–4122, Toronto, Canada. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, WWW '20, page 201–211, 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 P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2021. Revisiting the evaluation of end-to-end event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4609–4617, Online. Association for Computational Linguistics.

A Limitations

A.1 LLM Judgments *v.s.* Human Judgments

In this paper, advanced LLM GPT-4o is used to automatically judge the correctness of the predicted trigger words and arguments for the event extraction task. However, it should be noted that LLMs cannot always perfectly align with human judgments, which is the most reliable evaluation. For example, GPT-4o achieves comparable agreement and correlations with human average performance in ED and EAE annotations. Since the ED precision task is challenging, the Spearman correlation score on it is significantly lower than human performance. While we have made efforts to improve the relevance between LLM and human assessment through prompt design, this gap still exists. In the future, we would like to discover other strategies to further improve the reliability of LLM-as-a-Judge, such as multi-agent debate (Chan et al., 2023).

A.2 Computational Cost of Evaluating One Model

Unlike existing EM evaluation, our proposed metrics rely on advanced LLMs to annotate the quality of predicted events, leading to much higher costs. The average cost of evaluating one model on each benchmark is between 2-5$, which is smaller than established benchmarks like AlpacaEval (Li et al., 2023b).

A.3 Insufficient Evaluation for LLMs

The number of the test set in existing event extract datasets contains thousands of samples, leading to a huge cost for LLMs to do inference experiments. Following previous works (Huang et al., 2024), we sample a subset of the test set in existing datasets and prompt LLMs to handle the event

extraction task. However, we wish to emphasize that the subset consists of diverse samples from 10 event extraction datasets. Therefore, the conclusions of our experimental results are reliable and robust.

A.4 Other Advanced LLMs as Evaluator

In this paper, the GPT-4o model is used to conduct the automatic evaluation in RAEE. Due to the budget limitation, we did not employ a diverse set of advanced LLMs for this task, such as Claude-3 and PaLM. Since our preliminary study reveals that GPT-4o is slightly better than other advanced LLMs, we mainly use it as the automatic judge model in our proposed RAEE framework.

B Datasets Statistics

We utilize 10 widely-used datasets as evaluation benchmarks in both EM and RAEE evaluation. Since these datasets originate from various domains and differ in terms of pre-defined ontologies as well as size, it is necessary to introduce basic statistics of these datasets, as shown in Table 8.

Dataset	#Inst	#ET	#Evt	#RT	#Arg
ED and EAE tasks					
ACE05-en	20920	33	5348	22	8097
Genia2011	1375	9	13537	10	11865
CASIE	1483	5	8469	26	22575
PHEE	4827	2	5019	16	25760
ED task only					
FewEvent	12573	100	12573	—	—
MEE-en	13000	16	17257	—	—
M^2E^2	6013	8	1105	—	—
EAE task only					
WikiEvents	565	50	3932	58	5501
RAMS	9647	139	9647	65	21206
GENEVA	3684	115	7505	220	12314

Table 8: Statistics of 10 datasets used in RAEE. #Inst, #ET, #Evt, #RT, and #Arg represent the number of data instances, event type, event instances, role types, and argument instances, respectively.

C Key Criteria in RAEE

Considering that the requirements for EE vary across different scenarios and applications, we introduce key criteria in RAEE to enable the adaptive mechanism. As shown in Tabel 9, these key criteria are based on factors that significantly influence the evaluation results in our human annotation meta-evaluation experiment. Furthermore, this adaptive mechanism is easily extendable, allowing for the addition of personalized judgment criteria to account for more factors in extraction.

D Prompts

D.1 Reassessment Prompts in RAEE

Table 10, 11, 13 and 15 illustrate the prompts we use for reassessment in RAEE.

D.2 Inference Prompts for LLMs Extracting Events

Table 12 and 14 illustrate the prompts we use for testing the event extraction ability of LLMs in ED and EAE tasks, respectively.

E Implementation Details

To avoid inconsistency in evaluation caused by factors out of model frameworks such as data preprocessing or Python package versions, we follow TextEE (Huang et al., 2024), utilize its standardized data splits and unified input-output framework for evaluating models, and adhere to its default training configs, which can be found in the corresponding GitHub repository. All discriminative models are based on the RoBERTa-large (Liu et al., 2019) pre-trained model, and generative models are built using BART-large (Lewis et al., 2019).

Key Criteria Template	Setting
If there is a more accurate predefined type for a predicted trigger/argument, then only the more accurate type can be reassessed as { } one	correct
If a predict trigger/argument is more reasonable than golden annotations, it should be reassessed as { } one	correct
When there is co-reference, pronouns can be reassessed as { } one	correct
If the core word is correct but some modifiers are missing, the predicted trigger/argument should be reassessed as { } one	correct
A predicted trigger is reassessed as { } one when it triggers an event that does not really occur	incorrect
A golden trigger/argument is reassessed as { } one if there is no correspondence in predicted triggers/arguments, even if there is a predicted trigger/argument with a more accurate or reasonable type	not recalled
When there is co-reference, pronouns can be reassessed as { } one	recalled
If the core word is recalled but some modifiers are missing, the golden trigger/argument should be reassessed as { } one	recalled

Table 9: Key criteria templates and their settings used in RAEE.

Prompt Used for Reassessing on ED Task and Precision Metric	
Instruction	You are a reassessor designed to reassess predicted event triggers from an event detection model. Task Description: Refer to a piece of context and golden triggers, reassess predicted event triggers. A trigger is the key word in the sentence that most explicitly conveys the occurrence of the event. A predicted trigger can be reassessed as correct one if it conveys the occurrence of a particular interest kind of event, even if it does not exactly match any of golden triggers; otherwise, it is reassessed as incorrect one.
Key Criteria	There are several rules to follow when reassessing: (1) If there is a more accurate predefined type for a predicted trigger, then only the more accurate type can be reassessed as correct one; (2) If a predict trigger is more reasonable than golden annotations, it should be reassessed as correct one; (3) When there is co-reference, pronouns can be reassessed as correct one; (4) If the core word is correct but some modifiers are missing, the predicted trigger should be reassessed as correct one; (5) A predicted trigger is reassessed as incorrect one when it triggers an event that does not really occur.
Requirements	When answering, provide your thought process at first, including analyzing the context, explaining how golden triggers indicate events with a specific type, and reassessing predicted triggers one by one based on your thought. After that, answer final reassessment results. Please answer in the following python dict format: {'Thought Process': {'Context Analysis': str, 'Gold Triggers': [your explanation for gold trigger 0, your explanation for gold trigger 1, ...], 'Predicted Triggers': [your reassessment thought for predicted trigger 0, your reassessment thought for predicted trigger 1, ...]}, 'Final Reassessment Results': [1, 0, ...]}.
Data	Events of interest include: Contact.Meet. The event is related to contact meet. Business.Start-Org. The event is related to business start organization. Context: Professor Massimo Pigliucci , a [t.Pred.1] [t.Pred.0] professor [/t.Pred.0] [/t.Pred.1] of ecology and evolution at the State University of New York at Stony Brook , has criticized ICR for professing to present the same science as that taught in secular universities while at the same time requiring students and faculty to sign a statement of faith to ICR 's fundamentalist religious mission , most notably in affirming conformity in all its [t.Gold.0] work [/t.Gold.0] to Biblical doctrine . The position of each following trigger is labeled in the context between a pair of [t] and [/t]. Trigger Gold.0: work # event type: Business.Employment-Tenure Trigger Pred.0: professor # event type: Business.Employment-Tenure Trigger Pred.1: professor # event type: Education.Education

Table 10: Prompt used for reassessing on ED task and precision metric. An example of data is illustrated.

Prompt Used for Reassessing on ED Task and Recall Metric	
Instruction	You are a reassessor designed to reassess the recall result of an event detection model. Task Description: According to a piece of context, predicted triggers and golden triggers, reassess whether each golden trigger is recalled. A trigger is the key word in the sentence that most explicitly conveys the occurrence of the event. A golden trigger can be reassessed as recalled one if it can correspond to one or more predicted triggers that convey the occurrence of an event with the same interest type, even if it does not exactly match any of predicted triggers; otherwise, it is reassessed as not recalled one.
Key Criteria	There are several rules to follow when reassessing: (1) A golden trigger is reassessed as not recalled one if there is no correspondence in predicted triggers, even if there is a predicted trigger with more accurate or reasonable type; (2) When there is co-reference, pronouns can be reassessed as recalled one; (3) If the core word is recalled but some modifiers are missing, the golden trigger should be reassessed as recalled one.
Requirements	When answering, provide your thought process at first, including analyzing the context, explaining how golden triggers indicate events with a specific type, finding correspondence with predicted triggers for each golden trigger or explaining why any correspondence cannot be found, and reassessing golden triggers one by one based on your thought. After that, answer final reassessment results. Please answer in the following python dict format: {'Thought Process': {'Context Analysis': str, 'Gold Triggers': [your finding or explanation for gold trigger 0, your finding or explanation for gold trigger 1, ...]}, 'Final Reassessment Results': [1 , 0 , ...]}.
Data	Events of interest include: Business.Start-Subsidiary. The event is related to business start subsidiary. Context: A subsidiary company or daughter company is a company that is [t.Gold.0] owned or [t.Pred.0] controlled [/t.Pred.0] by [/t.Gold.0] another company , which is called the parent company , parent , or holding company . The position of each following trigger is labeled in the context between a pair of [t] and [/t]. Trigger Pred.0: controlled # event type: Business.Start-Subsidiary Trigger Gold.0: owned or controlled by # event type: Business.Start-Subsidiary

Table 11: Prompt used for reassessing on ED task and recall metric. An example of data is illustrated.

Prompt Used for Event Detection	
Instruction	You are an event extractor designed to check for the presence of a specific event in a piece of context and to locate the corresponding event trigger. Task Description: Identify all triggers related to the event of interest in the context. A trigger is the key word in the context that most explicitly conveys the occurrence of the event.
Requirements	When extracting, analyzing the context at first. After that, answer extraction results. You need to extract the span of a extracted trigger as well as its corresponding event type. Note that there may be zero to plural triggers in the context. Please answer in the following python dict format: {'Context Analysis': str, 'Extraction Results': [{'Trigger Span': a span in the context, 'Event Type': a specific event of interest}, ...]}.
Data	Types of events that may occur include: Adverse_event. The event is related to adverse event. Potential_therapeutic_event. The event is related to potential therapeutic event. Context: The case histories are presented of two patients who developed lung disease associated with the use of nitrofurantoin with histological features of bronchiolitis obliterans organising pneumonia (BOOP) , a rare but recognised form of drug induced injury .

Table 12: Prompt used for testing the event extraction ability of LLMs in ED task. An example of data is illustrated.

Prompt Used for Reassessing on EAE Task and Precision Metric	
Instruction	You are a reassessor designed to reassess predicted event arguments from an event argument extraction model. Task Description: Refer to a piece of context and golden arguments, reassess predicted event arguments. Arguments have the semantic role corresponding to the given event trigger by the word span between [t] and [/t]. A predicted argument can be reassessed as correct one if it have a particular interest type of role corresponding to the given event trigger, even if it does not exactly match any of golden arguments; otherwise, it is reassessed as incorrect one.
Key Criteria	There are several rules to follow when reassessing: (1) If there is a more accurate predefined role type for a predicted argument, then only the more accurate role type can be reassessed as correct one; (2) If a predict argument is more reasonable than golden annotations, it should be reassessed as correct one; (3) When there is co-reference, pronouns can be reassessed as correct one; (4) If the core word is correct but some modifiers are missing, the predicted argument should be reassessed as correct one.
Requirements	When answering, provide your thought process at first, including analyzing the context, explaining roles of golden arguments in the event of interest, and reassessing predicted arguments one by one based on your thought. After that, answer final reassessment results. Please answer in the following python dict format: {'Thought Process': {'Context Analysis': str, 'Gold Arguments': [your explanation for gold argument 0, your explanation for gold argument 1, ...], 'Predicted Arguments': [your reassessment thought for predicted argument 0, your reassessment thought for predicted argument 1, ...]}, 'Final Reassessment Results': [1, 0, ...]}.
Data	The event of interest is Justice:Trial-Hearing. The event is related to a trial or hearing for someone. Argument roles of interest: ['Defendant', 'Prosecutor', 'Place', 'Adjudicator'] Context: She was not [t] tried [/t] in the deaths of the other two . Gold argument 0: She # role type: Defendant Pred argument 0: deaths # role type: Defendant

Table 13: Prompt used for reassessing on EAE task and precision metric. An example of data is illustrated.

Prompt Used for Event Argument Extraction	
Instruction	You are an argument extractor designed to check for the presence of arguments regarding specific roles for an event in a piece of context. Task Description: Identify all event arguments related to roles of interest in the context. These arguments should have the semantic role corresponding to the given event trigger by the word span between [t] and [/t].
Requirements	When extracting, analyzing the context at first. After that, answer extraction results. Note that there may be zero to plural arguments for each role. Please answer in the following python dict format: {'Context Analysis': str, 'Extraction Results': {'role1': [str(the argument span), ...(The length of this list depends on the number of arguments that you extract for this kind of role)], 'role2': [str, ...], ...}}.
Data	The event of interest is justice.initiatejudicialprocess.trialhearing. The event is related to justice initiative judicial process trial hearing. Roles of interest include: prosecutor, defendant, judgecourt, crime, place. Context: But no longer . By the end of 2015 the establishment was expressing considerable dismay and desperation over its inability to do so , as the Republican base and its choices fell out of control . Republican elected officials and contenders for the next presidential election expressed open contempt for the Paris deliberations , refusing to even attend the proceedings . The three candidates who led in the polls at the time Donald Trump , Ted Cruz , and Ben Carson adopted the stand of the largely evangelical base : humans have no impact on global warming , if it is happening at all . The other candidates reject government action to deal with the matter .

Table 14: Prompt used for testing the event extraction ability of LLMs in EAE task. An example of data is illustrated.

Prompt Used for Reassessing on EAE Task and Recall Metric	
Instruction	You are a reassessor designed to reassess the recall result of an event argument extraction model. Task Description: According to a piece of context, predicted arguments and golden arguments, reassess whether each golden argument is recalled. Arguments have the semantic role corresponding to the given event trigger by the word span between [t] and [/t]. A golden argument can be reassessed as recalled one if it can correspond to one or more predicted arguments that have the same interest type of role corresponding to the given event trigger, even if it does not exactly match any of predicted arguments; otherwise, it is reassessed as not recalled one.
Key Criteria	There are several rules to follow when reassessing: (1) A golden argument is reassessed as not recalled one if there is no correspondence in predicted arguments, even if there is a predicted argument with more accurate or reasonable role type; (2) When there is co-reference, pronouns can be reassessed as recalled one; (3) If the core word is recalled but some modifiers are missing, the golden argument should be reassessed as recalled one.
Requirements	When reassessing, provide your thought process at first, including analyzing the context, explaining roles of golden arguments in the event of interest, finding correspondence with predicted arguments for each golden argument or explaining why any correspondence cannot be found, and reassessing golden arguments one by one based on your thought. After that, answer final reassessment results. Please answer in the following python dict format: {'Thought Process': {'Context Analysis': str, 'Gold Arguments': [your finding or explanation for gold argument 0, your finding or explanation for gold argument 1, ...]}, 'Final Reassessment Results': [1, 0, ...]}.
Data	The event of interest is Life:Die. The event is related to life and someone died. Argument roles of interest: ['Agent', 'Victim', 'Instrument', 'Place'] Context: There 's nothing very pretty about the training that you take to prepare you for combat , because it is to [t] kill [/t] people . Pred argument 0: people # role type: Victim Gold argument 0: you # role type: Agent

Table 15: Prompt used for reassessing on EAE task and recall metric. An example of data is illustrated.