MAVEN-FACT: A Large-scale Event Factuality Detection Dataset

Chunyang Li[1]*, **Hao Peng**[2]*, **Xiaozhi Wang**[2], **Yunjia Qi**[2],
Lei Hou[2], **Bin Xu**[2], **Juanzi Li**[2]†
[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
cliei@connect.ust.hk,
peng-h24@mails.tsinghua.edu.cn

Abstract

Event Factuality Detection (EFD) task determines the factuality of textual events, i.e., classifying whether an event is a fact, possibility, or impossibility, which is essential for faithfully understanding and utilizing event knowledge. However, due to the lack of high-quality large-scale data, event factuality detection is under-explored in event understanding research, which limits the development of EFD community. To address these issues and provide faithful event understanding, we introduce MAVEN-FACT, a large-scale and high-quality EFD dataset based on the MAVEN dataset. MAVEN-FACT includes factuality annotations of 112,276 events, making it the largest EFD dataset. Extensive experiments demonstrate that MAVEN-FACT is challenging for both conventional fine-tuned models and large language models (LLMs). Thanks to the comprehensive annotations of event arguments and relations in MAVEN, MAVEN-FACT also supports some further analyses and we find that adopting event arguments and relations helps in event factuality detection for fine-tuned models but does not benefit LLMs. Furthermore, we preliminarily study an application case of event factuality detection and find it helps in mitigating event-related hallucination in LLMs. Our dataset and codes can be obtained from https://github.com/THU-KEG/MAVEN-FACT

1 Introduction

Event Factuality Detection (EFD) aims to determine the factuality of textual events, i.e., classifying whether an event is a fact, possibility, or impossibility (Saurí and Pustejovsky, 2009, 2012; Lee et al., 2015; Veyseh et al., 2019; Murzaku et al., 2023). As shown in Figure 1, the event "*play*" is a fact while the event "*celebrate*" is just a possibility considering the word "*might*".

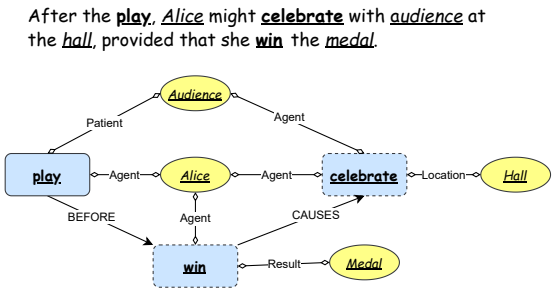

Figure 1: An example of event understanding. The event "*play*" is factual while the events "*win*" and "*celebrate*" are just possibilities considering the word "*might*".

Event factuality detection is a subfield of event understanding, which aims to extract structured event knowledge from plain texts (Wang et al., 2023a,b; Peng et al., 2023b; Huang et al., 2023a; Choudhary and Du, 2024), as shown in Figure 1. Event understanding is fundamental to broad downstream applications (Ding et al., 2015; Goldfarb-Tarrant et al., 2019; Wang et al., 2021). Previous event understanding work focuses on three primary tasks: event detection (Wang et al., 2020), event argument extraction (Wang et al., 2023a), and event relation extraction (Wang et al., 2022). However, event factuality detection is under-explored.

The primary reason for the under-exploration of EFD may be the lack of a large-scale, high-quality EFD dataset. Previous EFD datasets are usually small-scale. For example, the most widely-used dataset FactBank (Saurí and Pustejovsky, 2009) only includes 9,761 events, which may not provide sufficient data for model training and evaluation. Furthermore, these datasets also lack annotations for event arguments and relations, preventing a comprehensive understanding of events. In fact, considering factuality is crucial in event understanding. For example, if a downstream application mistakenly takes the "*celebrate*" event in Figure 1 as a fact rather than a possibility, it is likely to lead

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11140–11158

to erroneous judgments or even broader impacts.

To alleviate these issues, we introduce MAVEN-FACT, a large-scale and high-quality event factuality detection dataset based on MAVEN (Wang et al., 2020, 2022, 2023a). MAVEN provides a unified and comprehensive annotation, including event types (Wang et al., 2020), arguments (Wang et al., 2023a), and relations (Wang et al., 2022), for the same set of documents. Building on the solid foundation of the MAVEN series, this work extends the annotation to include event factuality. Therefore, MAVEN-FACT includes comprehensive annotations including event types, arguments, relations, and factuality, which can support faithful all-in-one event understanding. MAVEN-FACT also offers three main advantages: (1) **Large data scale**. MAVEN-FACT includes factuality annotations for 112,276 events, making it the largest EFD dataset. (2) **Supporting evidence annotation**. MAVEN-FACT also provides supporting evidence annotations, i.e., the words that directly convey factuality, e.g., *may*, for non-factual events. This enables a detailed analysis of factuality understanding of models and enhances models' interpretability by outputting supporting evidence of their factuality predictions (Zhao et al., 2024). (3) **Enabling task interaction**. Intuitively, some event information may help in factuality detection. For example, if an event has the *start-time* argument, then the event should be a fact. Thanks to MAVEN's event annotations, MAVEN-FACT enables analyzing how the event elements, including type, arguments, and relations, affect factuality detection, and vice versa.

To reduce cost and ensure data quality, we design an LLM-then-human annotation approach. Specifically, due to most events (exceeds 80%) being factual, we can endeavor to pre-annotate them automatically. We employ GPT-3.5 (OpenAI, 2022) for pre-annotation and formalize the task as a binary classification task (`factual` or `non-factual`) and develop a chain-of-thought prompt (Wei et al., 2022) method incorporating heuristic rules to ensure the high recall rate of the `non-factual` class. Subsequently, we manually annotate events pre-annotated as `non-factual`. To ensure data quality, the LLM pre-annotation is only used for the training set, while the events in validation and test sets are all human-annotated. This approach saves about 15% annotation costs (about 2,500 USD).

In the experiments, we evaluate several strong and representative models, including fine-tuned EFD models (Kenton and Toutanova, 2019; Liu et al., 2019; Wang et al., 2019; Murzaku et al., 2023), and LLMs with in-context learning (Brown et al., 2020), including Mistral 7B (Jiang et al., 2023), LLAMA 3 (Meta, 2024), GPT-3.5 (OpenAI, 2022), and GPT-4 (OpenAI, 2023). Experimental results demonstrate that the best-performing model only achieves a 47.6% macro F1 score and an even lower F1 score for non-factual events. It suggests that MAVEN-FACT is quite challenging for existing EFD models and LLMs. We conduct further experiments by requiring the models to provide supporting words for their predictions and find that this F1 score is much lower than that of EFD. It indicates that even if the model can correctly detect factuality, it may not provide accurate explanations. We also observe that adding arguments and relations enhances the performance of fine-tuned EFD models, whereas it does not benefit LLMs with in-context learning. Furthermore, we preliminarily study a potential application case of event factuality detection in mitigating event-related hallucination (Huang et al., 2023b), and find that incorporating event factuality can help mitigate hallucination in LLMs. We hope MAVEN-FACT and our empirical findings could facilitate future research on event factuality detection and faithful event understanding.

2 Dataset Construction

This section introduces the definition of event factuality detection (§ 2.1), the LLM-then-human annotation approach (§§ 2.2 to 2.4), and data analysis of MAVEN-FACT (§ 2.5).

2.1 Task Formulation

Event factuality detection is the task of assessing whether an event is a fact. Typically, this task is formalized as a multi-class classification problem. We adopt the widely-used 5 classes (Saurí and Pustejovsky, 2009; Qian et al., 2018a), *CT+*, *PS+*, *PS-*, *CT-*, and *Uu*, as the label set of MAVEN-FACT. These classes are based on the polarity and modality of event factuality, as illustrated in Figure 2. Modality indicates the certainty degree of events, where *CT* stands for certain and *PS* for possible. Polarity shows whether the event occurs or will occur, with + representing occurrence and - representing not occurring. To reduce annotation cost and bias, we do not adopt finer-grained class definitions like in Lee et al. (2015) and. These 5 classes are sufficient to express the polarity and modality for factuality detection and support its applications.

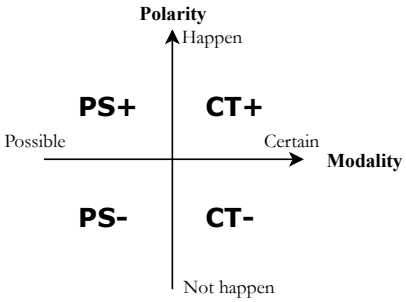

Figure 2: An illustration of four factuality classes. *Uu* denotes factuality can not be determined by the given context and is not shown in the figure.

The factuality source of events in MAVEN-FACT is the author because the author's belief in MAVEN objectively presents the event factuality. MAVEN-FACT also supports the supporting evidence prediction task (Alvarez Melis and Jaakkola, 2018), which predicts the words conveying the factuality. In this paper, we formalize this task as a pipeline: the models first perform EFD and then predict supporting words based on their factuality predictions.

2.2 LLM-then-Human Annotation Approach

In this paper, we aim to annotate the factuality of the overall 112,276 events from the MAVEN dataset to construct MAVEN-FACT. Due to the large scale of the data, manual annotation for all events is costly and not conveniently transferable to other domains or scenarios. Given the proven efficacy of LLMs as effective annotators (Mirzakhmedova et al., 2024; Chen et al., 2024), we develop an LLM-then-human annotation workflow to reduce costs while ensuring the annotation quality. We first adopt GPT-3.5 (OpenAI, 2022) to pre-annotate the data, filtering out events requiring human annotation, followed by a meticulous human annotation. This annotation approach reduces annotation costs by approximately 15%, saving about 2,500 USD. We only annotate the supporting words for *PS+*, *PS-*, and *CT-* events, as *CT+* and *Uu* events usually do not involve obvious supporting evidence. We employ this annotation workflow for the training set events only, while the validation and test sets are fully human-annotated. We finally sample 50 documents from the training set and find less than 2% noise, which demonstrates the efficacy of our annotation approach and the high quality of MAVEN-FACT. We will describe the details of LLM annotation (§ 2.3) and human annotation (§ 2.4) in the following sections.

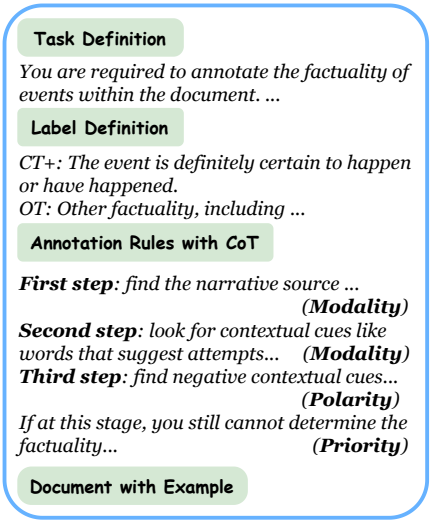

Figure 3: Prompt used in LLM pre-annotation.

2.3 LLM Annotation

Due to the majority of events being *CT+*, i.e., already occurred, we aim to employ LLMs to **automatically** distinguish between *CT+* and *non-CT+* events, and **only events pre-annotated as non-CT+ require further human annotation**, thereby reducing the need for human annotation. Consequently, the recall of pre-annotating *non-CT+* events is crucial for reducing pre-annotation noise. We improve the recall of *non-CT+* events in two main aspects: (1) We simplify the event factuality detection task into a binary classification problem, only distinguishing between *CT+* and *non-CT+* factuality, as binary classification is generally simpler than multi-classification (Rifkin and Klautau, 2004). Our objective is to obtain a high recall score for *non-CT+* to avoid filtering events that require further human annotation. (2) We adopt the chain-of-thought prompting method (CoT) (Wei et al., 2022) to better promote LLMs. Specifically, we first construct a comprehensive collection of annotation rules for the EFD task, integrated from multiple authors. Based on these rules, we design a step-by-step prompt that mirrors the human process for this task: ① Determine the narrative perspective of the event. ② Identify words conveying modalities, such as "may". ③ Check whether the text contains negations. ④ Default classification to *non-CT+* if factuality cannot be determined. Figure 3 illustrates the details of the prompt, and the full prompt can be found at appendix A.

To validate the efficacy of our annotation method, we conduct a pilot experiment on 500 human-

Setting	Direct Prompt		CoT Prompt	
	non-CT+	CT+	non-CT+	CT+
Multi-Class	18.8	95.7	55.2	69.2
Bi-Class	27.1	89.3	97.4	16.5

Table 1: The recall rate of *non-CT+* and *CT+* using direct and CoT prompts under multiple and binary classification settings. Higher *non-CT+* recall denotes less pre-annotation noise. Higher *CT+* recall means reducing more human annotation.

annotated events. We adopt GPT-3.5 as the LLM for pre-annotating. The experimental results are presented in Table 1. We can observe that our approach, binary classification with chain-of-thought prompting, achieves a 97.5% recall of *non-CT+*, which indicates it introduces little noise during pre-annotation, and a 16.5% recall of *CT+*, which suggests that the pre-annotation substantially reduce the annotation cost. Finally, we pre-annotate the overall training set, resulting in $16,950$ events labeled as *CT+* and $56,989$ as *non-CT+*, with the latter needing further human annotation.

2.4 Human Annotation

To ensure data quality, we manually annotate all events pre-labeled as *non-CT+* in the training set and all events in the validation and test sets. We employ a commercial annotation company for annotation, which involves 47 annotators, including 8 senior annotators responsible for annotation training and quality verification of other annotators's annotations. All annotations were performed on a specially developed platform. We ask the annotators to assess the factuality of events based solely on the provided text, without considering external knowledge. If annotators can not determine the factuality based on the provided text, they will label the event as *Uu*. For *PS+*, *PS-*, and *CT-* events, annotators are required to provide supporting evidence, which are the words extracted from the given text. If there are multiple supporting words, annotators are required to extract all of them. Senior annotators will randomly check 5% of annotated factuality. If an annotator's error rate assessed by the senior annotator exceeds 5%, they will undergo re-training for the annotation and all of this annotator's annotation will be re-labeled. We randomly sample 100 documents and annotate them twice by different annotator groups. The final inter-annotator agreement (accuracy) is 96.1%, demonstrating the annotation quality. Annotation details are provided in appendix A.

2.5 Data Analysis

Table 2 presents the statistics of MAVEN-FACT and other widely-used EFD datasets, including FactBank (Saurí and Pustejovsky, 2009), MEANTIME (Minard et al., 2016), UW (Lee et al., 2015), EB-DLEF (Zhang et al., 2022), DLFE-v2 (Qian et al., 2022), UDS-IH2 (Rudinger et al., 2018). More detailed data statistics of splitting MAVEN-FACT is provided in appendix A. We can observe that MAVEN-FACT possesses the largest data scale and includes supporting word annotations. Thanks to MAVEN's extensive annotations, MAVEN-FACT provides comprehensive annotations of events, arguments, relations, and factuality, supporting comprehensive and faithful event understanding research and applications.

3 Experiment

3.1 Experimental Setup

Baselines We evaluate several advanced and representative models, mainly including fine-tuned EFD models and large language models with in-context learning (Brown et al., 2020).

For fine-tuned EFD models, we reproduce several advanced models, including (1) **BERT+CLS** and **RoBERTa+CLS**, which adopt BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019) as the text encoder, respectively, and use the representation of a special token `[CLS]` (Kenton and Toutanova, 2019) as the factuality representation of the event for factuality classification. (2) **DMBERT** (Wang et al., 2019) and **DM-RoBERTa**, classical event understanding models that also utilize BERT and RoBERTa as the text encoder, respectively, and incorporate a dynamic multi-pooling mechanism (Chen et al., 2015) to integrate context and event information into a factuality representation for the final factuality classification. (3) **GenEFD** (Murzaku et al., 2023), a generative model based on FLAN-T5 (Chung et al., 2024). Murzaku et al. (2023) transform the event factuality detection task into a text generation task and design a meticulous factuality structure and target text structure, then optimize FLAN-T5 through multi-task learning. This model achieves state-of-the-art performance on FactBank. Our implementation employs the same setting except for not using multi-task learning as we only train the model on the event factuality detection task. For

Dataset	#Doc.	#CT+	#CT-	#PS+	#PS-	#Uu	#Total	Supporting Words
FactBank	208	7,749	433	589	70	4,619	13,460	✗
MEANTIME	120	1,798	44	83	3	125	2,053	✗
UW	276	–	–	–	–	–	13,923	✗
UDS-IH2	–	–	–	–	–	–	27,289	✗
EB-DLEF	7,840	5,222	1,601	935	53	29	7,840	✓
DLEF-v2	9,180	5,555	2,029	1,454	84	58	9,180	✓
MAVEN-FACT	4,480	105,209	2,330	3,874	540	323	112,276	✓

Table 2: Statistics of MAVEN-FACT compared with other event factuality detection datasets. Doc. is the short for Document. "Supporting Words" means whether the dataset contains supporting words of factuality. UW and UDS-IH2 adopt continuous factuality values and hence the statistics for discrete factuality are not applicable.

Model	CT+			CT-			PS+			PS-			Uu			Macro-F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
BERT+CLS	94.1	98.6	96.3	**66.6**	54.0	59.6	61.4	35.5	45.0	**61.0**	17.7	27.5	15.4	1.1	2.0	46.1
RoBERTa+CLS	94.1	98.6	96.3	61.0	54.8	57.8	61.3	32.0	42.0	44.8	19.2	26.9	**50.0**	2.2	4.1	45.4
DMBERT	94.4	98.4	96.3	64.8	55.9	60.0	62.2	37.5	46.8	45.6	23.2	30.7	26.7	2.2	3.1	**47.6**
DMRoBERTa	94.3	98.4	96.3	62.3	**60.4**	**61.3**	62.6	34.4	44.4	50.0	23.2	**31.6**	16.7	1.1	2.0	47.1
GenEFD	94.2	**98.7**	**96.4**	65.9	54.4	59.6	**63.8**	37.5	47.3	57.1	13.8	22.2	0.0	0.0	0.0	45.1
Mistral 7B	92.6	80.7	86.2	30.0	11.8	16.9	14.4	39.8	21.2	3.5	17.6	5.8	14.3	6.1	**8.5**	27.7
+CoT	93.0	70.3	80.1	9.0	17.6	11.9	11.6	38.3	17.8	2.6	23.5	4.8	10.5	6.1	7.7	24.4
LLAMA 3	91.7	75.3	82.7	18.2	11.8	14.3	10.2	38.3	16.1	0.0	0.0	0.0	7.7	6.1	6.8	24.0
+CoT	95.8	62.6	75.8	25.0	17.5	20.7	12.3	**71.9**	21.0	11.8	23.5	15.7	1.5	3.0	2.0	27.0
GPT-3.5	94.1	53.0	67.8	3.6	54.9	6.7	12.5	7.8	9.6	1.2	11.8	2.1	3.7	3.0	3.3	17.9
+CoT	**96.1**	27.9	43.2	4.1	36.3	7.3	8.4	47.7	14.3	3.1	**52.9**	5.8	3.3	**10.6**	5.1	15.1
GPT-4	94.1	94.6	94.4	51.4	37.3	43.2	44.7	56.3	49.8	16.7	11.8	13.8	0.0	0.0	0.0	40.2
+CoT	94.8	94.2	94.5	46.5	39.2	42.6	43.4	58.6	**49.8**	20.0	23.5	21.6	25.0	3.0	5.4	42.8

Table 3: Experimental results of fine-tuned EFD models and LLMs with in-context learning on MAVEN-FACT.

encoder-only models, we utilize cross-entropy loss for training. For training GenEFD, we employ language modeling loss (Bengio et al., 2000).

We also evaluate several LLMs with in-context learning, including two powerful open-sourced LLMs, **Mistral 7B** (Jiang et al., 2023) and the 8B version of **LLAMA 3** (Meta, 2024), and two proprietary LLMs, **GPT-3.5** (OpenAI, 2022) and **GPT-4** (OpenAI, 2023). For all experiments, we adopt 5-shot in-context learning. The demonstrations contain one exemplar from each category and are randomly sampled from the training set of MAVEN-FACT. We also evaluate LLMs with chain-of-thought prompt method (Wei et al., 2022), which is the same as in § 2.3 used for data annotation. Considering the time and monetary costs for LLMs inference, we sample 2,000 instances from the original test set of MAVEN-FACT to evaluate LLMs. More details of the experimental setup are placed in appendix B.

Evaluation Setup We adopt the same evaluation metrics with previous work (Qian et al., 2018a), and report precision (P), recall (R), F1 scores, and their macro averages for the *CT+*, *CT-*, *PS+*, and *PS-*, and *Uu* categories. For generative models, we use the exact match method (Rajpurkar et al., 2016) to compute the consistency rate between outputs and ground truth labels. For the chain-of-thought outputs, we require the LLM to provide its answer directly after "answer:" and automatically parse its response. If the outputs do not conform to this format, they are categorized as *Uu* predictions.

3.2 Experimental Results

The experimental results are shown in Table 3, and we have the following observations:

(1) Both the fine-tuned EFD models and LLMs exhibit moderate performance, particularly in the *CT-*, *PS+*, and *PS-* categories, compared to the results in the widely-used FactBank dataset (Qian et al., 2018a; Murzaku et al., 2023). This suggests that MAVEN-FACT poses a significant challenge to existing models. The small scale of existing datasets with limited *non-CT+* data may be insufficient for training and benchmarking EFD models, and hinders the development of advanced models. To further demonstrate that MAVEN-FACT is challenging and can serve as a valuable resource, we conducted a detailed error analysis. Additionally, we performed generalization experiments on FactBank. Both results are available in appendix B. We hope the large-scale MAVEN-FACT data will facilitate more research efforts to develop advanced mod-

els for the event factuality detection task. (2) LLMs significantly underperform fine-tuned models, especially in the *non-CT+* categories, and even the most powerful GPT-4 only achieves 42.8% macro F1. This aligns with previous findings that LLMs with in-context learning often fall short in information extraction tasks (Li et al., 2023; Han et al., 2023), possibly because LLMs lack specific understanding abilities to fine-grained information (Peng et al., 2023a), which is necessary for detecting factuality, such as "may". This suggests that LLMs may confuse event factuality, and we will show in § 4 that it results in non-factual responses of LLMs, i.e., hallucinations (Huang et al., 2023b), into the tasks requiring event knowledge. (3) The chain-of-thought approach has different effects on LLMs' performance. One possible reason is that the detailed instructions in the prompt enhance LLMs' fine-grained comprehension of texts, such as supporting words, while it may also cause overinterpretation of texts, leading to misclassification of *CT+* events into other categories. Although these results are still below those of fine-tuned models, it suggests that designing meticulous prompts to enhance the event factuality understanding ability of LLMs is feasible, and more research efforts are needed for enhancing this capability of LLMs, such as utilizing MAVEN-FACT as high-quality alignment data to align LLMs on the EFD task (Qi et al., 2024).

In conclusion, MAVEN-FACT presents a significant challenge to existing EFD models and LLMs. We hope that the high-quality MAVEN-FACT dataset will contribute to the training and benchmarking of EFD models and call for more research efforts to develop advanced EFD models.

3.3 Supporting Evidence Prediction

There are numerous works exploring explainability for models by requiring the models to provide explanations, i.e., supporting evidence, for their outputs, thereby enhancing the interpretability, transparency, and reliability of models (Luo and Specia, 2024). It is particularly essential for tasks involving factuality-related outputs where models are prone to generating hallucinations (Huang et al., 2023b). Therefore, for event factuality detection, providing coherent supporting evidence is essential for assessing the inherent understanding of event factuality and improving the reliability of models. However, as shown in Table 2, most previous datasets lack annotations for supporting evidence, i.e., the words conveying the factuality, leading to a lag in related research. MAVEN-FACT comprehensively provide annotated supporting words[1] for *CT-*, *PS+*, and *PS-* events to facilitate research on predicting supporting words and developing reliable EFD models.

Model	Factuality			Supporting Evidence		
	P	R	F1	P	R	F1
DMRoBERTa	74.5	**49.1**	**58.1**	**55.8**	39.4	**45.4**
GenEFD	**76.3**	40.5	50.4	49.5	**40.8**	44.7
LLAMA 3	53.7	14.3	18.5	4.6	2.8	3.5
GPT-4	62.4	32.6	42.5	21.0	18.3	19.5

Table 4: Macro averages of precision (P), recall (R), and F1 scores of *CT-*, *PS+*, *PS-* on the factuality and supporting evidence prediction task. We report the macro averages for only these three categories because only they have supporting evidence in the given input text.

We evaluate the EFD models on the task of predicting supporting words for their factuality predictions in the MAVEN-FACT dataset. Specifically, we employ a pipeline form where the model first detects the event factuality and then predicts supporting words based on its predicted factuality. Given that an event may have multiple supporting words, we use the sequence labeling paradigm (Akhundov et al., 2018) for models to predict these words. Without loss of generality, we evaluate four representative models, including DMRoBERTa, GenEFD, LLAMA 3, and GPT-4. Further experimental details are provided in appendix B. Table 4 presents the results, and we can find that the performance of supporting word prediction is significantly inferior to that of event factuality detection. This indicates that providing supporting words is more challenging, and models may struggle to provide valid supporting evidence even when they accurately predict factuality.

We further conduct an error analysis of supporting word prediction, and the errors stem from two main sources: incorrect factuality prediction and incorrect supporting word prediction. We categorize the errors into three types: `OnlyF`, `OnlyW`, and `Both`, which denote the errors come from only factuality prediction, only supporting word prediction, and both, respectively. The results are presented in Table 5. We can observe that (1) About 30% of the errors are `OnlyW`, indicating that even if the model accurately predicts factuality, it may still struggle to correctly predict the supporting words. (2) Except for LLAMA 3, a significant portion of

[1] In this paper, we refer to supporting evidence as negative or possibility cues provided as specific words or phrases, and there are no clear corresponding cues for *CT+* events.

Model	OnlyF	OnlyW	Both
DMRoBERTa	6.8	31.0	62.2
GenEFD	24.6	31.6	43.8
LLAMA 3	0.6	25.4	74.0
GPT-4	9.6	37.6	52.9

Table 5: Error rate (%) on supporting word prediction. `OnlyF`, `OnlyW`, and `Both` mean the errors come from only factuality prediction, only supporting word prediction, and both, respectively.

Model	Precision	Recall	Macro F1
DMRoBERTa	57.2	43.5	47.1
+relation	63.7	44.1	49.1
+argument	63.2	45.4	49.3
+both	53.5	41.2	45.6
GenEFD	56.2	40.9	45.1
+relation	57.7	41.4	45.4
+argument	54.4	43.0	46.4
+both	53.8	44.7	47.6
LLAMA 3	25.6	26.3	24.0
+relation	21.9	25.3	11.6
+argument	23.1	29.4	19.4
+both	23.5	25.1	16.9
GPT-4	41.4	40.0	40.2
+relation	42.4	35.9	37.9
+argument	41.4	36.0	37.6
+both	43.2	37.6	39.7

Table 6: Performance on event factuality detection after adding different event information.

the errors are `OnlyF`, suggesting that although the model does not predict factuality correctly, it accurately identifies supporting words. These errors suggest that the models can not sufficiently explain their own outputs. This hurts the reliability of the model in event factuality prediction. More efforts are needed to develop more reliable EFD models.

3.4 Analysis on Task Interaction

Thanks to MAVEN's comprehensive annotations, MAVEN-FACT also facilitates research about the interactions between event elements, such as arguments and relations, and event factuality, which is under-explored previously due to a lack of comprehensive data. In this paper, we primarily investigate whether event arguments and relations can help in event factuality detection. Intuitively, additional event information can benefit EFD. For example, if an event has a "*time*" argument referring to the past date, the modality of the event's factuality is more likely certain.

We conduct experiments on four representative models, DMRoBERTa, GenEFD, LLAMA 3, and GPT-4, using MAVEN-FACT to investigate whether adding event arguments and relations can help in EFD. For GenEFD, LLAMA 3 and GPT-4, we introduce additional information by transforming arguments and relations into natural language forms and placing them in the original text input. For DMRoBERTa, except for adding them in the text input, we introduce additional information by concatenating the representations of arguments and relations to the event factuality representation, and then use this concatenated representation to classify the factuality. The representations of arguments and relations are the average representations of their corresponding tokens. More experimental details can be found in appendix B.

The results are shown in Table 6, and we have the following observations: (1) For fine-tuned EFD models, DMRoBERTa and GenEFD, the experimental results generally align with our expectations, where the introduction of event arguments or relations tends to boost factuality detection performance. It suggests that fine-tuning models could better learn these correlations. However, adding both relation and argument information hurts the performance of DMRoBERTa. One possible reason is that the concatenated relation and argument representations may cause the model to more easily overfit to certain patterns in the training set. (2) For LLMs with in-context learning, introducing additional information tends to worsen performance. We find that the decline primarily comes from *CT+*, and the models are sensitive to prompts (Dong et al., 2022) and shifted towards classifying factuality as *non-CT+*. This suggests that few-shot in-context learning might introduce some biases instead of generalizable patterns (Si et al., 2023). More efforts are needed to effectively introduce additional information by in-context learning for event factuality detection, such as using many-shot in-context learning (Agarwal et al., 2024).

4 Mitigating Event-related Hallucinations

In addition to benchmarking EFD models, we also want to explore potential application scenarios of MAVEN-FACT. Here, we preliminarily explore using event factuality to mitigate event-related hallucinations in LLMs, as non-factuality is a primary source of hallucination (Huang et al., 2023b).

Hallucination refers to the phenomenon where the outputs of models do not align with the input, typically involving non-factual information in the

Document: The 2014 Bukidnon bus bombing occurred on December 9, 2014. ... Extortion is viewed as a motive for the attacks due to *claims* that the bus company has faced threats for refusing to *pay* protection money to the militants. The militant group denies any involvement claiming they would not gain any benefit from conducting such attacks and claims the accusations against them as fabrication.
Question: Did the bus bombing occur because the bus company refused to *pay* protection money to the militants?
Answer: No

Table 7: An instance from our constructed QA dataset. The "*pay*" event is a *PS-* event.

outputs (Huang et al., 2023b). This issue is prevalent in existing LLMs, raising concerns about the reliability and faithfulness of LLMs. There are numerous works exploring detecting and mitigating hallucination in LLMs (Ji et al., 2023; Dhuliawala et al., 2023; Yang et al., 2023; Zhang et al., 2024; Li et al., 2024). Event-related hallucination refers to the model outputting incorrect information about an event given its context, such as erroneous event arguments or causal relations, which is underexplored in previous research. Here, we hope to explore whether providing explicit event factuality information can help mitigate event-related hallucinations in LLMs.

Experimental Setup We begin with constructing a knowledge-intensive question-answering (QA) dataset based on MAVEN-FACT, which is a scenario susceptible to hallucinations (Huang et al., 2023b). As we aim to analyze hallucination in LLMs, we deliberately craft questions that are prone to induce hallucination. Specifically, we first sample 800 documents from the MAVEN-FACT test set. For each document, we select the event which is *non-factual* and have the most relation connections with other events in the document.[2] We then utilize GPT-4 to generate yes-or-no questions that require complex reasoning along with the answers for each event based on its mentioned document. Then three experts manually review all the questions and answers, and eliminate questions without answers or not requiring reasoning, and correct erroneous answers, resulting in 450 validated instances. An example of the data is shown in Table 7. We evaluate two representative LLMs, LLAMA 3 and GPT-4. For adding event factuality, we adopt two settings: (1) Oracle setting, which adds the ground truth factuality. This setting allows controlled experiments to observe the efficacy of adding factuality. (2) Real-world setting, which adds the DMRoBERTa predicted factuality and aligns with real-world scenarios. More experimental details are placed in appendix C.

[2]Having more relation connections suggests that the event involves more knowledge in the document, making it easier to construct knowledge-intensive questions about the event.

Setting	Factuality Info	LLAMA 3	GPT-4
Vanilla	✗	77.6	83.3
Real-World	✓	86.2	94.4
Oracle	✓	88.9	97.8

Table 8: Accuracy (%) on the constructed QA dataset. "Vanilla" denotes not adding factuality information.

Experimental Results The results are shown in Table 8. We can observe that adding factuality information (Oracle setting) significantly improves the accuracy of LLMs, i.e., reducing the hallucination rate. Using the factuality automatically extracted using DMRoBERTa is also effective. It offers a promising direction for research on reducing event-related hallucinations, namely by integrating additional factuality detection tools to explicitly include key information such as the factuality of events, triplets in the input, thereby mitigating model hallucinations. We encourage further exploration using MAVEN-FACT on this topic and to investigate more potential applications.

5 Related Work

5.1 Event Factuality Detection Datasets

FactBank (Saurí and Pustejovsky, 2009) is one of the earliest and widely-used EFD datasets. It is constructed based on TimeBank (Pustejovsky et al., 2006) and includes 5 types of factuality. MEANTIME (Minard et al., 2016) annotates a multilingual corpus of news articles with events and their corresponding factuality value, including certainty and polarity. To represent richer factuality, UW (Lee et al., 2015) and UDS-IH2 (Rudinger et al., 2018) adopted a continuous factuality value with a $[-3, 3]$ range. Recently, some studies introduced document-level EFD datasets, such as EB-DLEF (Zhang et al., 2022) and DLFE-v2 (Qian et al., 2022). MAVEN-FACT adopts 5 discrete factuality categories to enhance annotation quality and reduce subjective bias, which we believe sufficiently represents factuality.

5.2 Event Factuliaty Detection Methods

Conventional event factuality detection methods primarily use neural-based models, mainly including developing novel network architectures (Rudinger et al., 2018; Qian et al., 2018b; Veyseh et al., 2019; Cao et al., 2021; Liu et al., 2022) and designing new objectives (Qian et al., 2018a, 2019; Zhang et al., 2023). In the era of generative models, Murzaku et al. (2023) transformed the event factuality detection task into a text generation form, utilizing FLAN-T5 (Chung et al., 2024) for factuality detection. In this paper, we also evaluate LLMs and find that MAVEN-FACT poses significant challenges to existing methods.

6 Conclusion

This paper introduces MAVEN-FACT, the largest and high-quality event factuality detection dataset. MAVEN-FACT comprehensively includes supporting evidence for factuality and event annotations from MAVEN. Experimental results demonstrate that MAVEN-FACT poses a significant challenge to EFD models and LLMs. We also find that using event factuality can help in mitigating event-related hallucinations in LLMs. We hope that MAVEN-FACT will facilitate research on the development and application of event factuality detection.

Limitations

We discuss the limitations of this work here: (1) Language coverage. MAVEN-FACT only supports English, which may limit the widespread usage and application of our data. In the future, we will try to cover more languages and encourage community efforts for developing multilingual MAVEN-FACT. (2) Annotation approach. Our annotation approach only saves approximately 15% annotation cost. To ensure quality, we still employ substantial human annotation. However, this 15% reduction means a saving of about $2,500$ USD. We encourage the community to develop more advanced automated annotation methods using MAVEN-FACT. (3) LLM performance. We do not explore more prompting methods to enhance the performance of LLMs. We think this does not affect our experimental conclusions. LLMs typically underperform in specification-heavy tasks (Peng et al., 2023a) and require further efforts to improve their performance in EFD task.

Ethical Considerations

We discuss ethical concerns here: (1) **Intellectual property.** The MAVEN-ED dataset is released with CC BY-SA 4.0 license[3]. The MAVEN-ARG and MAVEN-ERE are published with GPLv3[4] license. We strictly adhere to their licenses when using these data. (2) **Intended use.** MAVEN-FACT is an event factuality detection dataset. Researchers and developers can use MAVEN-FACT to develop more advanced EFD methods and applications. (3) **Potential risk control.** MAVEN-FACT is constructed from public data, which we believe has been well anonymized and desensitized. The data annotation process does not include any personal or sensitive information of the annotators. We believe MAVEN-FACT introduces no additional risks. We will not release the test set and instead use an online scoring platform following previous work (Rajpurkar et al., 2016; Wang et al., 2020, 2022, 2023a) to prevent potential cheating use and data contamination (Xu et al., 2024), thereby ensuring a fair evaluation. (4) **AI assistance.** The writing of this paper employs ChatGPT to paraphrase some sentences.

Acknowledgement

We thank all the anonymous reviewers and meta reviewers for their valuable comments. This work is supported by Beijing Natural Science Foundation (L243006). This work is also supported by a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003).

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Stephanie Chan, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, et al. 2024. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

Adnan Akhundov, Dietrich Trautmann, and Georg Groh. 2018. Sequence labeling: A practical approach. *arXiv preprint arXiv:1808.03926*.

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of NeurIPs*.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Proceedings of NeurIPs*.

[3] https://creativecommons.org/licenses/by-sa/4.0/

[4] https://www.gnu.org/licenses/gpl-3.0.html

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPs*, pages 1877–1901.

Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao. 2021. Uncertain local-to-global networks for document-level event factuality identification. In *Proceedings of EMNLP*, pages 2636–2645.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of AAAI*, pages 17772–17780.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL*, pages 167–176.

Milind Choudhary and Xinya Du. 2024. Qaevent: Event extraction as question-answer pairs generation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1860–1873.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Proceedings of IJCAI*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of NAACL: Demonstrations*, pages 89–97.

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

Kuan-Hao Huang, I Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, Heng Ji, et al. 2023a. A reevaluation of event extraction: Past, present, and future challenges. *arXiv preprint arXiv:2311.09562*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of EMNLP*, pages 1643–1648.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. In *Proceedings of NeurIPs*.

Xiao Liu, Heyan Huang, and Yue Zhang. 2022. End-to-end event factuality prediction using directional labeled graph recurrent network. *Information Processing & Management*, 59(2):102836.

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. *arXiv preprint arXiv:1907.11692*.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*.

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

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke Van Erp, Anneleen Schoen, and Chantal Van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Proceedings of LREC*, pages 4417–4422.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? *arXiv preprint arXiv:2404.09696*.

John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. 2023. Towards generative event factuality prediction. In *Findings of ACL*, pages 701–715.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023a. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*.

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 ACL*, pages 9206–9227.

James Pustejovsky, Marc Verhagen, Roser Sauri, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2006. Timebank 1.2.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. Adelie: Aligning large language models on information extraction. *arXiv preprint arXiv:2405.05008*.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018a. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of IJCAI*, pages 4293–4300.

Zhong Qian, Peifeng Li, Guodong Zhou, and Qiaoming Zhu. 2018b. Event factuality identification via hybrid neural networks. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25*, pages 335–347. Springer.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-level event factuality identification via adversarial neural network. In *Proceedings of NAACL-HLT*, pages 2799–2809.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2022. Document-level event factuality identification via reinforced multi-granularity hierarchical attention networks. In *Proceedings of IJCAI*, pages 4338–4345.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Ryan Rifkin and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of NAACL-HLT*, pages 731–744.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational linguistics*, 38(2):261–299.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of ACL*, pages 11289–11310.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of ACL*, pages 4393–4399.

Shichao Wang, Xiangrui Cai, Hongbin Wang, and Xiaojie Yuan. 2021. Incorporating circumstances into narrative event prediction. In *Findings of EMNLP*, pages 4840–4849.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, et al. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of EMNLP*, pages 926–941.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, et al. 2023a. Maven-arg: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. *arXiv preprint arXiv:2311.09105*.

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 EMNLP*, pages 1652–1671.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019. HMEAE: Hierarchical modular event argument extraction. In *Proceedings of EMNLP-IJCNLP*, pages 5777–5783.

Xingyao Wang, Sha Li, and Heng Ji. 2023b. Code4struct: Code generation for few-shot event structure prediction. In *Proceedings of ACL*, pages 3640–3663.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPs*, pages 24824–24837.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Teaching large language models to refuse unknown questions. In *Proceedings of NAACL*.

Heng Zhang, Zhong Qian, Peifeng Li, and Xiaoxu Zhu. 2022. Evidence-based document-level event factuality identification. In *Pacific Rim International Conference on Artificial Intelligence*, pages 240–254. Springer.

Zihao Zhang, Zhong Qian, Xiaoxu Zhu, and Peifeng Li. 2023. Code: Contrastive learning method for document-level event factuality identification. In *International Conference on Database Systems for Advanced Applications*, pages 497–512. Springer.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Appendices

A Details on Data Construction

This section introduces details on annotation of MAVEN-FACT, including details of annotation instruction (appendix A.1), annotation coordination (appendix A.2), and data distribution (appendix A.3).

A.1 Annotation Instruction

The prompt used in the LLM annotation is shown in Table 9. Events pre-labeled as *non-CT+* in the training set and events in the validation and test sets are manually annotated. During the annotation process, we incorporated heuristic rules derived from contextual information and event relations from MAVEN-ERE (Wang et al., 2022) to guide the annotators. Some examples of the annotation rules can be seen in Table 12. For each factuality label, we provided specific examples and detailed explanations. This made it easier for the annotators to differentiate accurately. Additionally, we developed a new online annotation platform to support efficient and precise annotation, as shown in Figure 4.

A.2 Annotation Coordination

We engage annotators from a commercial data annotation company, comprising senior levels for annotation training and quality verification and others for data annotation. There are 47 annotators in total, among whom 55% are male and 45% are female. All annotators receive fair compensation, with their salaries and workloads agreed upon in advance. Employment is contract-based and adheres to local regulations. The total cost for annotation, including both the factuality and supporting evidence, as well as the development of annotation platforms, amounts to approximately $14,000$ USD. We explained how the data would be used and obtained consent.

A.3 Data Distribution of MAVEN-FACT

We follow the original spilt of MAVEN in (Wang et al., 2020). The statistics of splitting subsets are shown in Table 10.

B EFD Experimental Details

In this section, we introduce the implementation details regarding general details (appendix B.1) and task-specific details (appendix B.2). Additionally, we provide the results of a more detailed error analysis of EFD task on MAVEN-FACT (appendix B.3) and generalization experiments on FactBank (appendix B.4).

B.1 General Implementation Details

For fine-tuned EFD models, we train the models on our train set with a learning rate of $1e-5$ and a batch size of 16 over 10 epochs based on their checkpoint from HuggingFace. Table 11 shows the correspondence between models and the checkpoints we used for training. We insert special tokens (<e> and </e>) around the trigger words of events in the text to indicate their positions, which are also used as the basis for dynamic multi-pooling for DMBERT and DMRoBERTa.

For large language models with in-context learning, we use the official OpenAI to evaluate GPT-3.5 and GPT-4, with the decoding sampling temperature set to 0, and other parameters kept as default. We utilize the checkpoints from HuggingFace to evaluate LLAMA 3 and Mistral 7B. The checkpoints and API we used are also shown in Table 11.

All experiments are performed in a single run. We conduct experiments on Nvidia GeForce RTX 3090 GPUs, totaling approximately 200 GPU hours. For GPT-3.5 and GPT-4, we spend about 300 USD in total.

B.2 Task Specific Details

For the event factuality detection task, we conduct sentence-level training and testing, each data item is a sentence with its marked events. We conduct GenEFD experiments with the prefix "Event factuality prediction" for each data item. The prompt used in the in-context learning experiments of large language models is listed in Table 13.

For the supporting evidence prediction task, we approach it as a token classification task. The input consists of a list of words in the sentence and we insert special tokens showing its factuality around the trigger word of the event. The output is a list of the same length, with each element indicating the type of the corresponding word. In the output, 'O'

[5] https://huggingface.co/google-bert/bert-large-uncased

[6] https://huggingface.co/FacebookAI/roberta-large

[7] https://huggingface.co/google/flan-t5-base

[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[9] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Task Definition You are required to annotate the factuality of events within the provided document. The event trigger words are marked with '(**' and '**)' for your reference. Assign one of the following two labels to each event based on its context:
Label Definition CT+: The event is definitely certain to happen or have happened. OT: Other factuality, including but not limited to, certain not to happen or have happened, possible to happen or have happened, etc.
Annotation Rules with CoT 1. Firstly, observe the narrative source of the event. If the event originates from roles within the document, label it OT. 2. Then, look for contextual cues like words that suggest attempts, such as 'try to' or 'seek to', and words indicating possibility, such as 'may' or 'might'. These events will also be labeled OT. 3. Events with indicators in the context, such as words conveying negative cues like 'stop' or 'prevent', will also be labeled OT. 4.If at this stage, you still cannot determine the factual certainty of an event, prioritize labeling it as OT.
Document with Example {DOCUMENT} The output should be the same document with factuality label assigned behind each event trigger word. Do not output any other information. Example: Document: The company (**announced**) that it will (**launch**) a new product next month. Output: The company (**announced**)(CT+) that it will (**launch**)(OT) a new product next month.

Table 9: Prompt used for LLM annotation process. The "DOCUMENT" part varies depending on the data item.

Subset	#Doc.	#CT+	#CT-	#PS+	#PS-	#Uu	#Total
Train	2,913	69,782	1,492	2,262	285	118	73,939
Dev	710	16,868	384	456	52	20	17,780
Test	857	18,559	454	1,156	203	185	20,557

Table 10: Statistics of splitting MAVEN-FACT. Doc. is the short for Document.

Model	Checkpoint / API
BERT / DMBERT	`bert-large-uncased` [5]
RoBERTa / DMRoBERTa	`roberta-large`[6]
GenEFD	`flan-t5-base`[7]
LLAMA 3	`Meta-Llama-3-8B-Instruct`[8]
Mistral 7B	`Mistral-7B-Instruct-v0.2`[9]
GPT-3.5	`gpt-3.5-turbo`
GPT-4	`gpt-4`

Table 11: The correspondence between model and checkpoints or APIs.

represents other types, 'B' signifies the beginning of supporting words and 'I' indicates the interior of supporting words. This definition applies to the fine-tuned EFD models. For large language models, to enhance their understanding of the task, we provide prompts in addition to the input, as shown in Table 14.

For the task interaction, we utilize different approaches depending on the models and the tasks. In terms of tasks, for event relations, we mark triggers that have causal relations to the event to be classified with different special tokens in the sentences, and then concatenate them. For event arguments, we arrange them in a Type, Entity key-value pair format. Regarding the models, for DMRoBERTa, we use the processed event relations and event arguments as the input of its encoder, using the average representation of the token sequence as their representation. These representations are then concatenated with the original representation for classification. For GenEFD, we directly concatenate the processed event relations and event arguments with the original input as the model's input. For large language models, we incorporate explanations of event relations and arguments into the original event factuality detection prompt, followed by the processed event relations and arguments. The newly added prompt parts is shown in Table 15.

B.3 Error Analysis of EFD

We analyze the error cases. As shown in Table 16, we can observe that the errors are mainly due to the modality classification error, i.e., confusion between "possible" and "certain". This may be due to neglect of modality in either pre-training or post-training, which needs further exploration.

We also investigate more specific error types in modality classification (Table 17) and polarity classification (Table 18). We can observe that models tend to predict events as already having occurred, i.e., fact. This may lead to event-related hallucinations in the models' output.

Modality Rules 1. Modality is labeled based on the context. Events with a possible (*PS*) modality usually have obvious hint words in the context, such as words with a tentative meaning like "try to", "seek to", or words indicating possibility like "may", "might". 2. The Factuality of an event needs to consider its narrative source. We regard the document itself as the standard source. If the narrative source of an event is an argument in the document and the narration includes the argument's subjectivity, the modality is possible (*PS*).
Polarity Rules 1. Polarity is labeled based on the context. Events with a negative (-) modality usually have obvious hint words in the context, such as negative cues like "prevent", "can not".
Relation Rules 1. For the event set $\mathbb{B}$ containing all the events *B* with the relation *B CAUSES A*, if any event *B* in $\mathbb{B}$ has the factuality of *CT+*, then the factuality of event *A* is *CT+*. 2. For the event set $\mathbb{B}$ containing all the events *B* with the relation *B PRECONDITIONS A* relationship, if the factuality of event *A* is *CT-*, then the factuality of any event *B* in $\mathbb{B}$ is *CT-*.

Table 12: Some annotation rules for human annotation process. "Modality Rules", "Polarity Rules" and "Relation Rules" represent the rules for classifying modality, classifying polarity, and utilizing relations, respectively.

Figure 4: Screenshot for the annotation platform. The trigger word "siege" is selected for annotation, highlighted in yellow. Events related to it are highlighted in blue and green based on their relation type.

B.4 Generalization Results on FactBank

We have conducted generalization experiments on FactBank. The results further demonstrate MAVEN-FACT is of high quality and can serve as a valuable resource to the community.

FactBank consists of factuality from two sources: author source and non-author source. We divide FactBank into two subsets "AUTHOR" and "OTHER", whose source is author and non-author respectively. Each subset was further split into training, validation, and test sets in an $8:1:1$ ratio. Due to the label spaces of FactBank and MAVEN-FACT not being the same, we remap the labels in FactBank to align with the taxonomy used in MAVEN-FACT, as shown in Table 19.

We adopt DMRoBERTa as the EFD model and use two training methods. For the baseline, we train DMRoBERTa merely on the FactBank training set for ten epochs. For MAVEN-FACT augmentation, we first train DMRoBERTa on MAVEN-FACT for two epochs and then adapt the trained DMRoBERTa on FactBank training set for ten epochs. We use the FactBank validation set to select the best-performing checkpoint.

The results are shown in Table 20 and Table 21. We can observe that on AUTHOR set the MAVEN-FACT Augmented model performs much better, which demonstrates the high quality of MAVEN-FACT. Furthermore, the MAVEN-FACT Augmented model also performs better on OTHER set, even if MAVEN-FACT only contains factuality of author source.

Considering the large scale and high quality of MAVEN-FACT, we believe that MAVEN-FACT can serve as a significant resource in event factuality detection and event understanding, facilitating re-

INSTRUCTION:
You are an event factuality classifier. Please annotate the factuality of events within the text. The events are marked with '<e>' and '</e>'. Assign one of the following five labels to the event:
CT+: The event is certain to happen or have happened.
CT-: The event is certain not to happen or have happened.
PS+: The event is possible to happen or have happened.
PS-: The event is possible not to happen or have happened.
Uu: The event factuality is unknown.

RULES:
Here are some annotation Rules:
1. Firstly, observe the narrative source of the event. If the event originates from roles within the document instead of the document itself, label it 'PS+' or 'PS-'.
2. Then, look for contextual cues like words that suggest attempts, such as 'try to' or 'seek to', and words indicating possibility, such as 'may' or 'might'. These events will also be labeled 'PS+' or 'PS-'.
3. Events with indicators in the context, such as words conveying negative cues like 'stop' or 'prevent', will also be labeled 'PS-' or 'CT-'.

INPUT:
Here is the text you need to generate the label for, please do not output other information other than the label.
TEXT: President Herbert Hoover then <e>ordered</e> the Army to clear the marchers' campsite.
LABEL:

Table 13: An example of prompt for event factuality detection task. The RULES part is used for inference with chain-of-thought prompt. The INPUT part varies depending on the data item.

search in faithful event understanding.

C Mitigating Hallucination

We employed GPT-4 (`gpt-4o-2024-05-13`) to construct the QA dataset based on MAVEN-FACT. The construction process can be divided into two stages, and Table 22 displays the prompt templates used in each stage. Moreover, Table 23 presents the prompt information corresponding to three test configurations.

INSTRUCTION:
Task Description:
You are given a list of tokens representing a sentence containing an event. The event is marked with '<factuality>' and '</factuality>', where 'factuality' indicates the factuality of the event. The possible values for factuality are:
- CT+ (certainly happened)
- CT- (certainly did not happen)
- PS+ (possibly happened)
- PS- (possibly did not happen)

Your task is to generate an output list. Each element in the output list should correspond to an element in the token list. Use the following tags:
- 'O' for tokens that are not part of an evidential basis.
- 'B' for the beginning of an evidential basis.
- 'I' for the inside of an evidential basis.

Carefully analyze the input sentence and identify the event marked by '<factuality>' and '</factuality>'. Identify the evidential basis words that support the factuality of the event. Generate the output list with 'O', 'B', and 'I' tags according to the given rules.
Ensure your output matches the format and corresponds accurately to the input token list.

EXAMPLE:
For example:
Input: ['Webster', 's', 'confession', 'did', 'not', '<CT->', 'match', '</CT->', 'the', 'forensic', 'evidence', '.']
Output: ['O', 'O', 'O', 'B', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
In this case, the event is "match", its factuality is CT-, and the evidential basis is "did not".

INPUT:
Input: ['The', 'driver', 'applied', 'the', 'brakes', 'and', 'reversed', 'the', 'engine', ',', 'but', 'was', 'unable', 'to', '<CT->', 'stop', '</CT->', 'in', 'time', '.']
Output:

Table 14: An example of the prompt for supporting evidence prediction task. The INPUT part varies depending on the data item.

EXPLANATION:
In addition to the text, the event is accompanied by CAUSE relations, which is the <c>event</c> that causes the <e>event</e>, PRECONDITION relations, which is the event that must happen before the <e>event</e>. You can use these relations to help you determine the factuality of the event.

Argument information is also provided for the event. The arguments are the entities that are involved in the event. You can use the arguments to help you determine the factuality of the event.

RELATIONS:
Cause Relations:
Most of them were <c>car bombs</c> and most targeted infrastructure, especially the transport network.
At least twenty bombs <c>exploded</c> in the space of eighty minutes, most within a half hour period.
Precondition Relations:
The bombings were partly a response to the breakdown of talks between the IRA and the British government.

ARGUMENTS:
Arguments:TYPE: Agent; ENTITY: IRA. TYPE: Location; ENTITY: Belfast.

Table 15: An example of additional prompt for task interaction compared to event factuality detection task. The RELATIONS part and ARGUMENTS vary depending on the data item.

Model	Modality Only	Polarity Only	Both
DMRoBERTa	80.02	10.74	9.24
GenEFD	76.43	13.69	9.88
LLAMA 3	60.32	13.49	26.19
GPT-4	51.02	37.76	11.22

Table 16: Error rate (%) on Event Factuality Detection. Modality Only, Polarity Only, and Both mean the error from modality classification only, polarity classification only, and both accordingly.

Model	CT2O	O2CT
DMRoBERTa	8.95	91.05
GenEFD	3.77	96.23
LLAMA 3	30.28	69.72
GPT-4	8.20	91.80

Table 17: Error rate (%) on modality detection error. CT2O means the modality label is "CT" and the prediction is not "CT" and O2CT means the modality label is not "CT" and the prediction is "CT".

Model	P2N	N2P
DMRoBERTa	15.63	84.37
GenEFD	5.24	94.76
LLAMA 3	0.0	100.0
GPT-4	18.75	81.25

Table 18: Error rate (%) on polarity detection error. P2N means the polarity label is "+" and the prediction is "-" and N2P means the polarity label is "-" and the prediction is "+".

MAVEN-FACT	FactBank
CT+	*CT+*
CT-	*CT-*
PS+	*PS+, PR+*
PS-	*PS-, PR-*
Uu	*Uu*

Table 19: Labels map between MAVEN-FACT and FactBank. MAVEN-FACT and FactBank represent the labels used in MAVEN-FACT and FactBank accordingly.

Settings	CT+	CT-	PS+	PS-	Uu	Macro-F1
Baseline	95.4	76.6	83.7	0.0	92.0	69.5
MAVEN-FACT Aug	95.1	76.2	88.4	100.0	92.1	90.4

Table 20: F1 scores (%) on AUTHOR set of FactBank. MAVEN-FACT Aug and Baseline represent DMRoBERTa model with and without MAVEN-FACT augmentation, respectively.

Settings	CT+	CT-	PS+	PS-	Uu	Macro-F1
Baseline	89.6	66.7	74.7	33.3	64.2	65.7
MAVEN-FACT Aug	91.2	78.6	73.7	33.3	62.2	68.2

Table 21: F1 scores (%) on OTHER set of FactBank. MAVEN-FACT Aug and Baseline represent DMRoBERTa model with and without MAVEN-FACT augmentation, respectively.

STEP1: Constructing Reasoning Chain

Please generate an incorrect reasoning chain containing the "{trigger}" event (marked with <e> and </e>) based on the following document.
Whenever possible, use the "{trigger}" event as the start or middle node event of the reasoning chain, rather than as the conclusion event. Note that the "trigger" event may not occur in the document, but in the reasoning chain, the "{trigger}" event must occur for the reasoning to be valid.
This chain of reasoning should try not to include events not mentioned in the document.
Please give the chain of reasoning in numerical order and the reasoning chain within 6 steps.
Document: {document}
Reasoning Chain:

STEP2: Constructing Question

Please give a question based on the above chain of reasoning. It should not be too simple or too difficult.
The question should satisfy the following conditions: If the question is answered based on the above chain of reasoning, the answer will be Yes. However, if the question is answered based on the fact that "{trigger}" in the chain of reasoning does not necessarily occur, the answer will be No. Please directly output the questions that meet the requirements and do not output others.

Table 22: Prompt template for constructing QA dataset based on MAVEN-FACT. The input and output of STEP1 are attached to the input of STEP2 as history. In real application, {trigger} and {document} are filled with their corresponding input entries.

Vanilla Setting

Please answer the questions according to the document below. Please answer Yes or No directly and do not enter other words.
Document: {document}
Question: {question}
Answer:

Real-World and Oracle Setting

Please answer the questions according to the document below.
Please carefully distinguish which events actually occurred in the document and which events are just possible events. Answer the questions based on what exactly happened in the document.
Please answer Yes or No directly and do not enter other words.
Document: {document}
Question: {question}
Note that the "{trigger}" event in the above document is not an exact occurrence, but a {factuality}.
Answer:

Table 23: Prompt template for Vanilla, Real-World, and Oracle Setting in § 4. In application, {document}, {question}, {trigger}, and {factuality} are populated with their respective inputs. Depending on the trigger's event factuality, {factuality} is assigned accordingly: "probable occurrence" for *PS*+, "probable non-occurrence" for *PS*-, and "definite non-occurrence" for *CT*-.