Improving Event Definition Following For Zero-Shot Event Detection

Zefan Cai[*†◇], **Po-Nien Kung**[*‡], **Ashima Suvarna**[‡], **Mingyu Derek Ma**[‡], **Hritik Bansal**[‡], **Baobao Chang**[◇], **P. Jeffrey Brantingham**[‡], **Wei Wang**[‡], **Nanyun Peng**[‡]
[†]University of Wisconsin - Madison, [‡]University of California, Los Angeles,
[◇]Peking University
zefncai@gmail.com, ponienkung@cs.ucla.edu

Abstract

Existing approaches on zero-shot event detection usually train models on datasets annotated with known event types, and prompt them with unseen event definitions. These approaches yield sporadic successes, yet generally fall short of expectations. In this work, we aim to improve zero-shot event detection by training models to better follow event definitions. We hypothesize that a diverse set of event types and definitions are the key for models to learn to follow event definitions while existing event extraction datasets focus on annotating many high-quality examples for a few event types. To verify our hypothesis, we construct an automatically generated Diverse Event Definition (DivED) dataset and conduct comparative studies. Our experiments reveal that a large number of event types (200) and diverse event definitions can significantly boost event extraction performance; on the other hand, the performance does not scale with over ten examples per event type. Beyond scaling, we incorporate event ontology information and hard-negative samples during training, further boosting the performance. Based on these findings, we fine-tuned a LLaMA-2-7B model on our DivED dataset, yielding performance that surpasses SOTA large language models like GPT-3.5 across three open benchmarks on zero-shot event detection. Our code and data can be found at https://github.com/PlusLabNLP/ZeroED

1 Introduction

Event detection (ED) focuses on identifying event triggers of specific event types in a given text with predefined event ontology. Prior work has studied event detection largely in a fully-supervised fashion (Wadden et al., 2019; Lin et al., 2020; Nguyen and Grishman, 2015; Nguyen et al., 2016; Han et al., 2019; Du and Cardie, 2020; Cheng et al., 2023b; Cai et al., 2023; Cheng et al., 2023a; Huang et al., 2020; Huang and Peng, 2020; **?**; Paolini et al., 2021; Ma et al., 2023b). While these work show promising performance on seen events, it cannot generalize well to long-tailed and unseen events (Ma et al., 2024; Zhang et al., 2022). To further enable generalization to low-resource events, prior work proposed to tackle few-shot event detection by training model on generated pseudo data (Ma et al., 2024; Kumar et al., 2020; Schick and Schütze, 2021). Despite the success in data-efficient event detection, they cannot zero-shot extract unseen events in real-time due to the need for prior training, limiting their applicability to a wider range of scenarios.

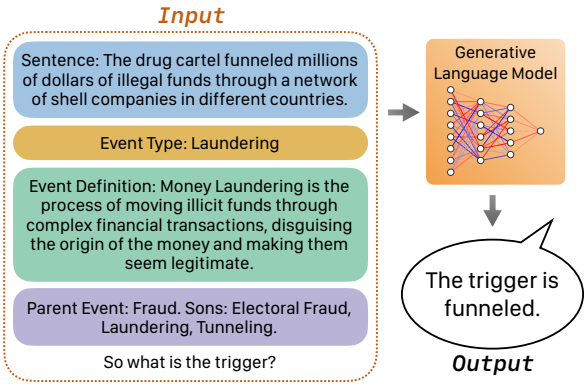

Figure 1: Zero-shot generative event detection formulation. We demonstrate a generated event type and sample from our DivED dataset. The input prompt includes information about *Event Type*, *Event Definition*, *Event Ontology* and the query passage, and the expected output is a verbalized extracted result.

The success of task generalization of LLMs enabled by instruction tuning further advances zero-shot event detection. Recent work started to extract events of novel type by providing LLMs with the event definition of unseen events during inference, as demonstrated in Figure 1. They either prompt closed-source LLMs, such as GPT-3.5 (Wang et al., 2022; Gao et al., 2023a; Wei et al., 2023), or apply transfer learning on

[*]Equal contribution.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2842–2863

open-sourced LMs with EE training data of seen event types (Huang et al., 2018a; Lyu et al., 2021). While the former methods achieve acceptable performance, they are not flexible and reproducible due to their closed nature, leading to the difficulty in further improving the models' performance. In contrast, the latter methods while reproducible and flexible, suffer from low performance.

In this work, we aim to enhance zero-shot event detection (ED) by training a model with improved generalization to unseen event types. During inference, the model, prompted with definitions of previously unseen events, relies on its instruction-following ability to understand event definitions and identify correct triggers. Despite recent impressive results of instruction following by Large Language Models (LLMs), there is room for improvement (Kung and Peng, 2023; Shao et al., 2023; Yin et al., 2023), and we focus on enhancing it via instruction fine-tuning with strategically generated data. Specifically, we hypothesize that transfer learning from conventional EE datasets might not be ideal (Kung et al., 2021, 2023). Though a large amount of high-quality training samples for only a few event types equips the model to perform EE on homogeneous data, it is not sufficient for the model to develop generalizability towards unseen situations. Different from existing works that aim to improve ED model by generating more homogeneous EE data (Ma et al., 2024), we posit that a large number of event types and a diverse set of event definitions are the keys to improving the event definition following capabilities.

To verify our assumption, we develop **Diverse Event Definition (DivED)** Dataset, which is generated from LLMs with diverse event definitions and samples. DivED includes 3000+ event types, each with 10 event definitions and 10 samples. Since event types can be organized into tree-structure ontology, we further inject each event's event type dependencies information into its event definition, including the name of its parent and sibling events.

Our study on the DivED dataset supports our hypothesis. The results indicate that a sufficient amount of event types (200) and diverse event definitions significantly enhance zero-shot event detection performance on out-of-distribution data, underscoring their crucial role in event definition comprehension. On the other hand, the performance doesn't improve significantly with more than ten samples per event type. This is attributed to the reliance of zero-shot event detection on the model's ability to generalize to new event types and definitions. While a few samples aid in learning the meaning of event types, an excessive number is unnecessary. In addition to scaling, we explore the impact of incorporating event ontology information in event definition and utilizing hard-negative samples during training. We observe that incorporating both components enhances the model's comprehension of event boundaries, resulting in higher recall and F1 scores.

Following this finding, we further train our model on the DivED and Geneva (Parekh et al., 2023a) datasets, and achieve state-of-the-art zero-shot event detection performance on ACE (Doddington et al., 2004a), M2E2 (Li et al., 2020a) and MEE (Veyseh et al., 2022) test sets benchmarked in TextEE (Huang et al., 2023), surpassing strong LLMs such as Chat-GPT with less than 5 percents of model parameters, showing the effectiveness and efficiency of our method. To conclude, our main contributions are as follows:

1. We design a data generation pipeline to generate a **Diverse Event Definition dataset (DivED)** with 3000+ event types and 10 diverse event definitions for each type. Our experiments reveal that diverse event types and event definitions are crucial to improve zero-shot ED.

2. We systematically study the **impact of various components of EE training data** on the ability of large instruction-tuned models to follow event definitions.

3. Our **proposed model achieves SOTA results** on ACE, M2E2, and MEE datasets, surpassing GPT-3.5-Turbo model with drastically fewer parameters.

2 Method

In this section, we describe our data generation pipeline to generate the *Diverse Event Definition Dataset (DivED)* and our systematic study on the impact of various components within event detection training data. We investigate how (1) the scaling of event types, event definition, and training samples and (2) incorporating ontology information and hard-negative samples can impact models' generalization to unseen event types.

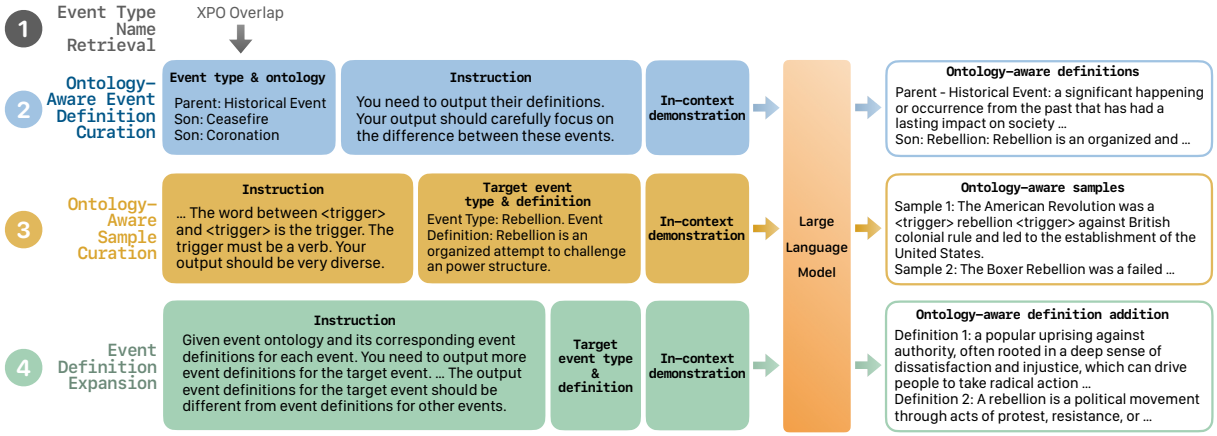

Figure 2: Data generation pipeline to generate DivED dataset. The pipeline includes five main steps: (1) Event Type Name Retrieval: retrieve events from XPO overlap (Spaulding et al., 2023b); (2) Ontology-Aware Event Definition Curation: generate event type definitions for the event types retrieved from (1); (3) Ontology-Aware Sample Curation: generate samples for the retrieved event type names from (1) and event definition from (2); (4) Event Definition Expansion: Paraphrase and expand the event definition from (2), and (5) Ontology Pruning: Prune out events with high trigger overlap. Details of our prompt templates can be found in Appendix B.

2.1 DivED Dataset Generation

An event detection dataset with diverse event types and definitions is necessary to investigate the effect of training data components systematically. Thus, we propose an automatic data generation pipeline that leverages proprietary LLM to generate *Diverse Event Definition Dataset (DivED)*. The data generation pipeline includes (1) Event Type Name Retrieval, (2) Ontology-Aware Event Definition Curation, (3) Ontology-Aware Sample Curation, (4) Event Definition Expansion, and (5) Ontology Pruning. Figure 2 illustrates the five-step data generation pipeline. We use ChatGPT (GPT-3.5-Turbo) for the dataset generation process. The average number of sibling event types is 2.14, and the number of ancestor event types is 1510. The total number of events is 3232. The examples of DivED and the data generation pipeline templates can be found in Appendix B.

Step 1: Event Type Name Retrieval We follow (Zhan et al., 2023) methods to collect around 6000 event type names with ontology (dependency trees) from XPO-overlap (Spaulding et al., 2023a), which provides a large set of event entities that occurred in Wikidata.* To guarantee the testing events from ACE (Doddington et al., 2004b), M2E2 (Li et al., 2020b), and (Veyseh et al., 2022) datasets are held out for our later experiments, we manually filtered out all events that share the same dependency trees with these testing events.

*wikidata.org

Step 2: Ontology-Aware Event Definition Curation After acquiring event type names and ontology (dependency trees of events), we instruct the model to simultaneously generate concise definitions for all event types within the ontology. Using one manually curated in-context example, we guide the model to differentiate similar events within the same ontology, resulting in distinct and well-distinguished event definitions, as demonstrated in Table 1.

Step 3: Ontology-Aware Sample Curation We follow a similar method as in **Ontology-Aware Event Definition Curation** to prompt the model with relative event types, event definition, and one manually curated in-context example to generate ten samples for multiple event types simultaneously. Each generated sample includes an input sentence and an output trigger of the corresponding event type. The generated samples can be seen in Table 1.

Step 4: Event Definition Expansion To get multiple event definitions for each event type, we prompt the model to expand or paraphrase the event definition ten times with the provided event type name, event definition, event ontology, and one manually curated in-context example.

Step 5: Ontology Pruning After generating data for all event types, we further prune out duplicate events within the same event ontology by identifying their output trigger overlap. Specifically, for an event ontology tree $\{\mathbf{e}_1, \mathbf{e}_2, ...\} \in \mathbf{E}$ with multiple event types and ten samples per event, we

Event	Event Definition	Sample	Trigger
Arriving	**Event Definition 1:** The act of Arriving involves the physical or virtual arrival at a destination ... **Event Definition 10:** The Arrival event captures the moment when someone ...	**Sample 1:** The school field trip participants **arrived** at the museum and were greeted ... **Sample 10:** The visitors **arrived** at the aquarium and were led to the dolphin show by the staff.	arrived, ...
Drop in on	**Event Definition 1:** Drop in on refers to an unplanned and impromptu visit to a friend or acquaintance ... **Event Definition 10:** The act of drop in on signifies an unscheduled visit to an individual's place ...	**Sample 1:** Renee decided to **pop in** on her friend who lived nearby and catch up. ... **Sample 10:** Jane had some free time on her hands and wanted to **pay a visit** to her former college roommate who lived close by.	pop in, pay a visit, ...
Visiting scenario arrival	**Event Definition 1:** Visiting scenario arrival entails arriving at a planned destination ... **Event Definition 10:** The event of Visiting scenario arrival involves arriving ...	**Sample 1:** The investors **arrived** at the company's headquarters for their business presentation. ... **Sample 10:** The family **reached** the theme park with their pre-booked ride tickets.	arrived, reached, ...

Table 1: We demonstrate the generated event definition and samples of a few sibling event types in the DivED dataset. During data curation, we specifically prompt models to generate distinct event type definitions and samples for these similar event types to enhance the diversity of the generated data.

calculate the output trigger overlap ratio between two event types $\mathbf{e}_i, \mathbf{e}_j$ where $i \neq j$. The trigger overlap is measured by exact string matching each of the ten triggers in $\mathbf{e}_i$ with the ten triggers in $\mathbf{e}_j$. If the overlap ratio of output triggers exceeds a certain threshold (in our implementation, it is 0.5), we will consider one of the two events as duplicate and remove it from our dataset. This way, we can guarantee that the event types and output triggers of our dataset are diverse.

2.2 Data Impact Analysis

With the generated **DivED** dataset, we systematically study the impact of various components in training data to understand how to train the model with improved event definition following ability.

Scaling of data components In Figure 1, we show the data components within the training data. During training, we will provide several samples, each corresponding to an event type and definition, to query the models about the event trigger. In testing time, we will further test on the unseen events, in which all the event types, definitions, and samples are unseen from training. This requires the model to generalize to the unseen events to be able to perform well during testing. Following this intuition, we aim to investigate how different numbers of events, definitions, and samples can influence models' performance. Specifically, for each dataset component, we fix the quantity of other components and evaluate the scaling law associated with it. For example, to investigate the scaling of event definition, we will use different number of event definitions per event, with a fixed amount of event type and samples.

Ontology information We further look into the construction of event definition and negative samples. In most zero-shot EE methods, they solely provide the information (event type, event definition) of the current event, without providing the event ontology information. We explore adding ontology information to the input definition in order to see how it helps models with the understanding of the event, and generalize better to unseen event types. The additional ontology information includes the parent and child events of an event ontology, as shown in Figure 1.

Hard-negative samples During model training, we use input sentences paired with output triggers. Positive samples are based on ground truth events, where the output for event trigger identification is not "None." To incorporate predictions for "None" events, we create negative samples by prompting the model with input sentences and an event type that does not occur in the sentence. In this work, we aim to explore how integrating ontology information into negative sample construction affects the model's ability to learn event definitions and boundaries. Instead of randomly assigning unrelated events during negative sample creation, we will assign sibling event types to form **hard-negative samples**. These challenging examples may offer additional signals about event boundaries that aid the model in distinguishing between similar events and improve its understanding of event definitions.

3 Experimental Setup

In this section, we first describe the details of **Data Impact Analysis** experiments, which analyze the impact of different data components. We further describe baselines and training details in **Enhancing**

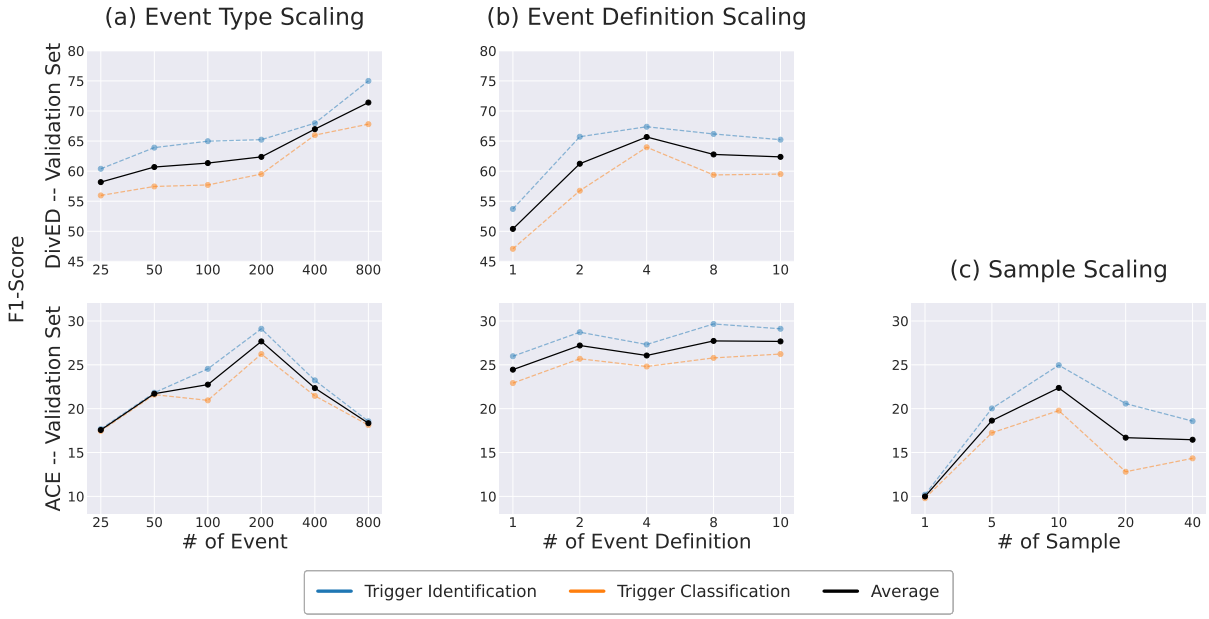

Figure 3: The scaling of different dataset components. We train the models with different number of event types, event definitions per event type and samples per event type. After training, we further report the F1 scores on *DivED – Validation* and *ACE Validation set*. Note that we do not report the *DivED – Validation* score separately for sample scaling as we utilize the Geneva (Parekh et al., 2023a) train set to explore sample scaling rather than DivED train set.

Zero-Shot Event Detection, in which we integrate the optimal settings from Data Impact Analysis to train our zero-shot event detection model and compare to previous state-of-the-art large language models (LLMs) on three event extraction datasets.

3.1 Data Impact Analysis

Model and Training Details We utilize LLaMA-2-7B (Touvron et al., 2023) for our experiments to examine the impact of various data components. Employing a batch size of 96, a learning rate of 2e-5, and training for 20 epochs. We divide the DivED dataset into training and validation sets by random sampling event ontologies. For the training set, We consistently use the DivED dataset with unified variables: 200 Event Types, 10 Event Definitions, 10 Samples, and 10 negative samples per sample. For the validation set, we first random sample around 50 event ontology, each including a parent event and multiple child events, in total 100 event types. The validation set can be considered an in-domain zero-shot ED evaluation set with all test events unseen from training but in a similar domain (all generated by ChatGPT), while ACE is a out-of-domain zero-shot ED test set previously curated by human annotators and has many different data characteristics, as described in Table 2. In each scaling experiment, only one variable is scaled while others are fixed. Notably, ACE-related events are excluded from the DivED dataset to ensure the test events are entirely novel.

Event type and event definition scaling We experiment on [200, 400, 800, 1600, 3200] events on the DivED dataset for event-type scaling. For Event definition scaling, we test [1, 2, 4, 8, 10] event definition to investigate the scaling law of these variables.

We employ the DivED dataset for definition and event type scaling due to its extensive collection of event definitions and types, a feature absent in traditional event detection datasets like the Geneva dataset. We present the evaluation results of the DivED validation set for event type and definition scaling since it reflects in-domain performance, as both DivED-train and DivED-val are generated using the LLM with similar prompts.

Sample scaling For sample scaling, since DivED only has ten samples per event, we conduct it on the Geneva (Parekh et al., 2023a) dataset and test with [1, 5, 10, 20, 40] samples per event. We filter out all ACE-related events for the Geneva dataset to make sure the test events are unseen.

We have studied sampling scaling using the GENEVA dataset, containing up to 40 high-quality human-generated event samples, a resource lacking in our DivED datasets (which have ten samples per event). When using the Geneva dataset, evaluation

results on the DivED validation set do not represent in-domain performance like event type and definition scaling, which may cause confusion to the reader. Thus, we do not show the DivED-val result here. While there exist prior work studying sample scaling, our focus is to study event definition and types scaling since the feature is absent in traditional ED datasets.

From sample scaling experiments on the Geneva dataset, the main insight is that a sufficient amount (10) of samples are sufficient for zero-shot event detection, and we further follow this insight to generate the DivED dataset. It is possible that the DivED data samples can suffer from the diversity issue, and the exact number of samples to achieve optimal performance can be different from Geneva; however, we show that using ten samples can already achieve performance superior to proprietary LLM such as ChatGPT. Further hyperparameter tuning and optimization might achieve higher performance but is not the focus of our paper.

Event Ontology and Hard Negative Samples In event ontology experiments, we assess two settings: with or without event ontology. For hard negative samples, we experiment using zero or three hard negative samples within the ten negative samples. Evaluation is conducted on the ACE dataset.

Evaluation We report the F1 scores of **event trigger identification** and **event trigger classification** on DivED (in-domain) and ACE (out-domain) validation set. In in-domain evaluation, 50 unseen events from DivED (absent in training) are tested. For out-domain assessment, the model is tested on the ACE dataset, comprising 33 event types unseen during training. We analyze the models' zero-shot generalization on these sets, presenting a comparison in Figure 3. Notably, DivED (in-domain) exhibits similar event definition and sample length to the training set, while ACE (out-domain) has longer definitions and a different writing style, focusing on argument details alongside the event itself.

3.2 Enhancing Zero-Shot Event Detection

Training Details Following our observation in **Data Impact Analysis**, we further train the LLaMA-2-7B (Touvron et al., 2023) models on DivED and Geneva (Parekh et al., 2023a) dataset following the optimal setting. We aim to show that our method has the potential to achieve performance surpassing large language models such as ChatGPT with a small 7B LLaMA model. By mixing the event type definitions from DivED and GENEVA, we can further improve the diversity of the input and event definition styles to achieve improved performance, surpassing ChatGPT models with orders of magnitude fewer model parameters. By merging these two datasets, we also demonstrate that our method is scalable and has the potential to be combined with existing EE datasets to achieve improvements. We use 200 + 90 event types from DivED and Geneva (Parekh et al., 2023a) datasets. We use ten event definitions, ten samples, and ten negative samples per sample for each event type while incorporating the ontology information and three hard-negative samples.

DivED-Train – Event: Money Laundering
Event Definition: *Money Laundering is the process of moving illicit funds through complex financial transactions, disguising the origin of the money and making them seem legitimate.* **Avg. Def. Length**: 42.9; **Avg. Sample Length**: 23.3
DivED-Validation – Event: Ceasefire
Event Definition: *A ceasefire is a mutual agreement between opposing armed groups to halt all aggressive actions and refrain from initiating any new hostilities, often negotiated to allow for the delivery of aid and the creation of safe zones for civilians.* **Avg. Def. Length**: 42.2; **Avg. Sample Length**: 23.3
Ace-Validation – Event: BE-BORN
Event Definition: *BE-BORN Event occurs whenever a PERSON Entity is given birth to. Please note that we do not include the birth of other things or ideas.* **Avg. Def. Length**: 65.3; **Avg. Sample Length**: 35.6

Table 2: Dataset comparison. We show the comparison of the definition, average definition token length and average sample token length between **DivED train set**, **DivED validation set** and **ACE test set**.

Baselines In our experiments, we conduct a comparison between our finetuned LLaMA-2-7B (Touvron et al., 2023) and several zero-shot event detection baselines, including ChatGPT (OpenAI, 2021), ChatIE (Wei et al., 2023) and (Gao et al., 2023a) and LLaMA-2-Geneva. Prompt templates for the baselines are provided at Appendix C.

- **ChatGPT (OpenAI, 2021)**: GPT-3.5-Turbo was prompted with the proposed method for a fair comparison with finetuned LLaMA-2-7B.
- **ChatIE (Wei et al., 2023)**: ChatIE is a framework that transforms the zero-shot event detection task into a multi-turn question-answering problem. Here, LLMs are first prompted (as shown in Appendix C) to identify the event type and then sequentially prompted to iden-

tify the trigger.

- **Gao et al. (2023a)**: This work explores the feasibility of ChatGPT as a zero-shot event detection model and further analyses the impact of event definitions, in-context examples and counterfactual examples in the prompt template presented at Appendix C. We prompt ChatGPT with event definitions and positive examples in our implementation as this setup performed best on Gao et al. (2023a) evaluation.

- **LLaMA-2-Geneva**: We additionally train an LLaMA2-7B model (Touvron et al., 2023) on Geneva (Parekh et al., 2023a) datasets as a transfer learning baseline. We first filter out all events related to ACE, M2E2, and MEE datasets from the training set, leaving 90 event types. We further train the model with all samples on the remaining event types.

Evaluation Datasets Our experiments comprehensively compare our fine-tuned LLaMa-2-7B with baselines across three popular event extraction benchmarks, including ACE05, M2E2, and MEE. We consider the English annotations of these datasets and report the F1 scores of event trigger identification and event trigger classification on their test set process by TextEE (Huang et al., 2023).

- ACE05 (Doddington et al., 2004b) is an end-to-end event extraction dataset which covers texts from several sources such as newswire, broadcast news and weblogs.

- M2E2 (Li et al., 2020b) is an end-to-end event extraction dataset collected from the multimedia domain. We only consider the text part.

- MEE (Veyseh et al., 2022) is a multilingual end-to-end event extraction dataset collected from Wikipedia which is extended from MINION (Song et al., 2015).

4 Results

4.1 Data Impact Analysis

Scaling of event types In Figure 3 column (a), we show the results of training a LLaMA-2-7B model with different numbers of events. It is seen that scaling up the number of events consistently helps the model performance on the in-domain DivED validation set. However, while training the model with more events continuously scales up the performance on the ACE validation set under 200 events, using more than 200 events leads to degeneration of the performance. This can be caused by the model overfitting to the domain of training data. While we continuously train on new events, the model can still overfit to the domain of the data itself, for example, the format of the event definition and samples. This also shows that while our generated **DivED** dataset has a large number of events, the generated samples and event definition might still have spurious correlations that can lead to overfitting.

Metric →	Trigger ID			Trigger CLS		
Model ↓	Prec.	Rec.	F1	Prec.	Rec.	F1
Ours	45.3	22.0	29.1	36.6	20.4	26.2
w/o Ontology	55.0	11.8	19.4	34.3	11.1	16.7
w/o Hard Neg.	53.5	16.6	25.3	45.5	15.6	23.2

Table 3: We report the experiment results of providing ontology information and using hard negative samples.

Scaling of event definition In Figure 3 column (b), we show the results of event definition scaling. On both the DivED and ACE validation sets, the performance scales up with four event definitions per event. While using more than four event definitions does not help the in-domain performance, it can help the model generalize better to the out-domain test set. This shows that adding more diverse event definitions during training can further improve the model's robustness. Helping the model to generalize to more diverse event formats during inference, thus improving zero-shot performance.

Scaling of samples To evaluate how using more samples helps the model's zero-shot generalization, we experiment using the Geneva dataset, which has a large number of high-quality samples per event type, and test on the ACE validation set. The results are shown in Figure 3 column (c). Surprisingly, while using more samples usually helps models' performance in a supervised setting, using more than ten samples hurts models' performance. This can be caused by the model overfitting the training data and becoming less robust to unseen events.

Following the results above, we conclude that the key to improving models' zero-shot generalization to unseen events is to use a diverse set of event definitions with a certain amount of event types and samples. While a small amount of event type and samples helps, using too much

	ACE						M2E2						MEE						Average	
Metric →	Trig. ID			Trig. CLS			Trig. ID			Trig. CLS			Trig. ID			Trig. CLS			Trig. ID	Trig. CLS
Model ↓	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	F1	F1
ChatIE	4.8	11.9	6.8	2.5	6.4	3.6	4.3	31.4	7.5	2.6	19.2	4.6	12.4	31.2	17.2	7.5	24.2	11.9	10.5	6.7
Gao et al.	42.5	28.7	34.2	30.6	20.6	24.6	16.6	38.0	23.1	14.5	33.2	20.2	84.7	7.8	14.2	77.9	7.15	13.9	23.8	19.6
GPT-3.5	9.2	60.7	15.9	3.6	43.9	6.6	7.0	63.3	12.6	4.9	54.2	8.9	12.5	33.2	18.2	7.6	24.4	11.6	15.6	9.0
Geneva	47.7	14.5	22.2	18.2	13.8	15.7	19.1	17.9	18.5	17.9	17.0	17.4	70.5	24.5	**36.4**	63.0	24.1	**34.9**	25.7	22.6
Ours	46.7	26.9	**34.2**	36.7	24.1	**29.1**	21.2	26.1	**23.4**	19.8	24.7	**22.0**	70.9	16.7	27.1	65.7	16.2	26.0	**28.2**	**25.7**

Table 4: The experiment results on ACE ,M2E2 and MEE test set. We compare the performance of LLaMA-2-7b training on DivED dataset (Ours) with ChatIE (Wei et al., 2023), Gao et al. (2023a), GPT3.5 model and LLaMA-2-7b trained on Geneva (Parekh et al., 2023a) dataset. We report the Precision, Recall and F1 scores. We also report the average F1 score across all datasets. We abbreviate Trigger as Trig.

can make the models overfit to the training source, leading to a degeneration of the generalization ability. This effect can be specifically obvious in machine-generated data, which can have spurious correlations and lack diversity in certain aspects.

Event ontology and hard-negative samples In Table 3, we further investigate the usefulness of the event ontology and hard-negative samples. It can be seen that after removing the event ontology or hard-negative samples, fewer triggers are predicted, which leads to a much lower recall and F1 score. This means that the model becomes more conservative at predicting triggers. We hypothesize that the model can be trained to distinguish similar events by providing ontology and hard negative samples. At testing time, this can help the models be more certain at predicting the triggers and increase the number of matching triggers.

4.2 Enhancing Zero-Shot Event Detection

Following the observation from **Data Impact Analysis**, we further apply the best setting and compare it with baseline models described in section 3.2. We show the results in Table 4. It can be seen that LLaMA-2-7B trained on the DivED dataset (Ours) consistently outperforms GPT baselines (ChatIE, Gao et al. (2023a) and GPT-3.5) on all ACE, M2E2, and MEE datasets and surpasses our LLaMA2-Geneva baselines on ACE and M2E2 datasets. For the MEE dataset, LLaMA2-Geneva achieves the best performance. Upon further investigation into the predicted results, we found that LLaMA2-Geneva can better predict samples that have multiple ground truth event triggers in one event type, which frequently occurred in MEE Geneva datasets but less occurred in ACE, M2E2 and DivED datasets, directly leading to higher Recall and F1 scores on MEE dataset. Generally, our proposed model achieves the best average F1 scores on both Event Trigger Identification and Event Trigger Classification, showing the superiority of the training method.

Metric →	Trigger ID			Trigger CLS		
Model ↓	Prec.	Rec.	F1	Prec.	Rec.	F1
Geneva	47.7	14.5	22.2	18.2	13.8	15.7
DivED	43.5	21.87	29.1	36.6	20.43	26.23
Div. + Gen.	46.7	26.9	34.2	36.7	24.1	29.1

Table 5: Ablation study of training data.

5 Discussion

5.1 Ablation Study of Training Data

As mentioned in section 3.2, we train the LLaMA-2-7B (Touvron et al., 2023) models on DivED and Geneva (Parekh et al., 2023a) dataset. In Table 5, we further provide results that only use the DivED dataset without GENEVA.

It can be seen that, by solely training on the DivED dataset, we can already achieve significant performance improvement and surpass model training on the human-curated Geneva dataset. When combined with the Geneva dataset, we can further boost the performance, showing the good scalability of our method.

For the results in Table 5, it does show that the current models are still limited to generalize to expert domains without domain adaptation. This can be due to the models' lack of the basic knowledge to solve these expert domains. While our work provides a method to improve event definition following for better zero-shot ED, the base model we used (llama-2) does not have information about these expert domains and still requires further domain adaptation to achieve a generally good performance on them. This limitation also happens with low-resourced language. If the base model is not trained on a specific knowledge, even if our curated dataset helps with event definition following, the model cannot perform it well, due to the

Metric → Model ↓	Trigger ID F1	Δ	Trigger CLS F1	Δ
Ours	29.1		26.2	
w/o Def	14.15	-52%	10.83	-59%
Geneva	25.2		17.4	
w/o Def	23.44	-7%	8.2	-53%

Table 6: We assess model performance drop by removing event definitions during training. We compare LLaMA-2 models trained on Geneva and DivED datasets. A higher drop rate indicates greater reliance on event definitions.

lack of language information. Note that while our proposed DivED dataset does not directly generalize to these expert domains or other languages, the insight from data impact analysis can be utilized to generate specific datasets for these domains, and potentially help the zero-shot ED performance.

5.2 Do Models Follow the Event Definition?

Instruction-tuned models excel in various zero-shot tasks but can excessively rely on the spurious patterns within the provided prompt, neglecting instruction semantics (Kung and Peng, 2023; Yin et al., 2023). In this work, we aim to enhance zero-shot event detection by training model with better event definition following. To assess the model's event definition following ability, we conduct an ablation study following prior work's setting (Kung and Peng, 2023), comparing our LLaMA-2-7B model trained on DivED data with one trained on the conventional EE dataset Geneva (Parekh et al., 2023b). To verify whether our proposed model have better event definition following ability compare to models learning on convention EE datasets, we follow prior work (Kung and Peng, 2023) to conduct an ablation study. We compare our LLaMA-2-7B (Touvron et al., 2023) model trained on DivED data with a LLaMA-2-7B model trained on conventional EE dataset such as Geneva (Parekh et al., 2023b). Despite having numerous samples per event, conventional EE dataset has only one definition per event type, which largely differs from DivED dataset. We report the performance drop rate after removing the event definition during training and testing for both models in Table Table 6. It can be seen that while the performance drops for both models after removing the event definition during training, the model trained on the DivED dataset has a higher performance drop, especially in *Event Trigger Identification*, showing that our proposed model heavily relies on event definition during training and inference. This indicates that our model is better at utilizing the event definition information, potentially exhibiting a better event definition following ability.

6 Related Work

Low-resource information extraction Low-resource IE models (Si et al., 2023; Hu et al., 2023) secure their performance with limited training data by cross-task transfer learning that uses supervision from tasks like Semantic Role Labeling (Zhang et al., 2021; Huang et al., 2018b), indirect supervision that reformulates the task as data-rich tasks like NLI or QA (Xu et al., 2023; Sainz et al., 2022; Ma et al., 2023a; Lu et al., 2022), both heavily rely on task compatibility. Some works focus on prompting generative LMs with enriched task requirements and examples (Li et al., 2023; Gao et al., 2023b), which is constrained the diversity of human-curated training data.

Data generation for IE Existing works explore different strategies to generate training data instances given a known task output space through analogous input (Kumar et al., 2020; Lee et al., 2021), creating pesudo labels with weak annotator (He et al., 2021; Chia et al., 2022; Ye et al., 2022; Wang et al., 2023; Schick and Schütze, 2021), reverse generation (Meng et al., 2022; Gao et al., 2021; Josifoski et al., 2023) and structure-to-text generation (Ma et al., 2024). Different from introducing more data instances for observed task space, we instead aim to extend the model's generalizability by generating new types and their definitions for unseen data distribution that extend the task space with LLM-oriented data generation.

7 Conclusion

We investigate how incorporating diverse event types and definitions benefits zero-shot event detection models. The proposed DivED dataset features a large number of diverse event types and definitions, which helps train the model to better generalize to unseen event definitions. By further incorporating event ontology and hard negative samples, we finetuned a LLaMA-2 model on DivED and Geneva datasets, which consistently surpasses previous SOTA ChatGPT prompting baselines in zero-shot ED on ACE, M2E2, and MEE datasets. Overall, our findings provide insights to improve models' event definition following ability and provide an opportunity to further advance zero-shot ED on open-sourced models.

Limitation

Our study on zero-shot event detection, despite its advances, faces several limitations. The reliance on automatically generated datasets may not fully capture complex real-world events, potentially limiting the model's generalizability. Additionally, the effectiveness of our approach depends on detailed event ontology and the availability of hard-negative samples, which might not always be accessible. Scalability also poses a challenge, as expanding the diversity of event types requires significant computational and data resources. Moreover, our findings are primarily based on English language benchmarks, raising questions about the applicability of our results across different languages and domains. Future research should address these limitations to enhance the robustness and universality of zero-shot event detection models.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback. This work was partially supported by AFOSR MURI via Grant #FA9550-22-1-0380, Defense Advanced Research Project Agency (DARPA) grant #HR00112290103/HR0011260656, CISCO and ONR grant #N00014-23-1-2780.

References

Zefan Cai, Xin Zheng, Tianyu Liu, Xu Wang, Haoran Meng, Jiaqi Han, Gang Yuan, Binghuai Lin, Baobao Chang, and Yunbo Cao. 2023. Dialogvcs: Robust natural language understanding in dialogue system upgrade.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023a. Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Proc. of EMNLP*.

Xuxin Cheng, Zhihong Zhu, Wanshi Xu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023b. Accelerating multiple intent detection and slot filling via targeted knowledge distillation. In *Proc. of EMNLP*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. RelationPrompt: Leveraging Prompts to Generate Synthetic Data for Zero-Shot Relation Triplet Extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004a. 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).

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004b. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. *arXiv preprint arXiv:2004.13625*.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023a. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023b. Exploring the Feasibility of ChatGPT for Event Extraction.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Rujun Han, I Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, Nanyun Peng, et al. 2019. Deep structured neural network for event temporal relation extraction. *arXiv preprint arXiv:1909.10094*.

Xuanli He, Islam Nassar, Jamie Ryan Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2021. Generate, Annotate, and Learn: Generative Models Advance Self-Training and Knowledge Distillation.

Helan Hu, Shuzheng Si, Haozhe Zhao, Shuang Zeng, Kaikai An, Zefan Cai, and Baobao Chang. 2023. Distantly-supervised named entity recognition with uncertainty-aware teacher learning and student-student collaborative learning.

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

Kung-Hsiang Huang and Nanyun Peng. 2020. Document-level event extraction with efficient end-to-end learning of cross-event dependencies. *arXiv preprint arXiv:2010.12787*.

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. *arXiv preprint arXiv:2009.09335*.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018a. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018b. Zero-Shot Transfer Learning for Event Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data Augmentation using Pre-trained Transformer Models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks.

Po-Nien Kung, Sheng-Siang Yin, Yi-Cheng Chen, Tse-Hsuan Yang, and Yun-Nung Chen. 2021. Efficient multi-task auxiliary learning: Selecting auxiliary data by feature similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 416–428, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *arXiv preprint 2102.01335*.

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.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020a. 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.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020b. Cross-media structured common space for multimedia event extraction. *arXiv preprint arXiv:2005.02472*.

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.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332.

Mingyu Derek Ma, Jiun-Yu Kao, Shuyang Gao, Arpit Gupta, Di Jin, Tagyoung Chung, and Nanyun Peng. 2023a. Parameter-Efficient Low-Resource Dialogue State Tracking by Prompt Tuning. In *INTERSPEECH 2023*.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023b. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. Association for Computational Linguistics.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. Star: Improving low-resource information extraction by structure-to-text data generation with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 300–309.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371.

OpenAI. 2021. ChatGPT: Large-scale language model. Accessed: June 17, 2023.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023a. 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.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023b. 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.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual Entailment for Event Argument Extraction: Zero- and Few-Shot with Multi-Source Learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Timo Schick and Hinrich Schütze. 2021. Generating Datasets with Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Nan Shao, Zefan Cai, Hanwei xu, Chonghua Liao, Yanan Zheng, and Zhilin Yang. 2023. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.

Shuzheng Si, Zefan Cai, Shuang Zeng, Guoqiang Feng, Jiaxing Lin, and Baobao Chang. 2023. Santa: Separate strategies for inaccurate and incomplete annotation noise in distantly-supervised named entity recognition.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Elizabeth Spaulding, Kathryn Conger, Anatole Gershman, Rosario Uceda-Sosa, Susan Windisch Brown, James Pustejovsky, Peter Anick, and Martha Palmer. 2023a. The darpa wikidata overlay: Wikidata as an ontology for natural language processing. In *Workshop on Interoperable Semantic Annotation (ISA-19)*, page 1.

Elizabeth Spaulding, Gary Kazantsev, and Mark Dredze. 2023b. Joint end-to-end semantic proto-role labeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 723–736, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Mee: A novel multilingual event extraction dataset. *arXiv preprint arXiv:2211.05955*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Xingyao Wang, Sha Li, and Heng Ji. 2022. Code4struct: Code generation for few-shot event structure prediction. *arXiv preprint arXiv:2210.12810*.

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

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2023. Can NLI Provide Proper Indirect Supervision for Low-resource Biomedical Relation Extraction? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. *arXiv preprint arXiv:2306.01150*.

Qiusi Zhan, Sha Li, Kathryn Conger, Martha Palmer, Heng Ji, and Jiawei Han. 2023. Glen: General-purpose event detection for thousands of types. *arXiv preprint arXiv:2303.09093*.

Hongming Zhang, Haoyu Wang, and Dan Roth. 2021. Zero-shot Label-Aware Event Trigger and Argument Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Hongming Zhang, Wenlin Yao, and Dong Yu. 2022. Efficient zero-shot event extraction with context-definition alignment. *arXiv preprint arXiv:2211.05156*.

A Cost Estimates for OpenAI

We implement all baselines in Section 3.2 with GPT-3.5-Turbo. The estimated cost of implementing our baselines is approximately 100 USD. Similarly, the estimated cost of implementing our baselines on GPT-4 will be approximately 3000 USD and we leave this implementation for future due to limited resources. This further emphasizes that our finetuned model surpasses larger LLMs in performance as well as accessibility due to the cost effectiveness of the method.

For Event Definition Curation, Sample Curation, and Event Definition Expansion, the expense of using GPT models was less than 100 USD; for Event Type Name Retrieval and Ontology Pruning, the whole process was automatically conducted without labor or expenses. Therefore, the effort needed to produce the curated dataset is very small compared to manually collected datasets.

B Data Generation of DivED dataset

Step 1: Event Type Name Retrieval We follow (Zhan et al., 2023) methods to collect around 6000 event type names with ontology (dependency trees) from XPO-overlap (Spaulding et al., 2023a), which provides a large set of event entities that occurred in Wikidata.[†] To guarantee the testing events from ACE (Doddington et al., 2004b), M2E2 (Li et al., 2020b), and (Veyseh et al., 2022) datasets are held out for our later experiments, we manually filtered out all events that share the same dependency trees with these testing events.

Step 2: Ontology-Aware Event Definition Curation After acquiring event type names and ontology (dependency trees of events), we instruct the model to simultaneously generate concise definitions for all event types within the ontology. Using one manually curated in-context example, we guide the model to differentiate similar events within the same ontology, resulting in distinct and well-distinguished event definitions, as demonstrated in Table 1. The template utilized in this step is presented at Table 7.

Step 3: Ontology-Aware Sample Curation We follow a similar method as in **Ontology-Aware Event Definition Curation** to prompt the model with relative event types, event definition, and one manually curated in-context example to generate

[†] wikidata.org

ten samples for multiple event types simultaneously. Each generated sample includes an input sentence and an output trigger of the corresponding event type. The generated samples can be seen in Table 1. The template utilized in this step is presented at Table 8.

Step 4: Event Definition Expansion To get multiple event definitions for each event type, we prompt the model to expand or paraphrase the event definition ten times with the provided event type name, event definition, event ontology, and one manually curated in-context example. The template utilized in this step is presented at Table 9.

Step 5: Ontology Pruning After generating data for all event types, we further prune out duplicate events within the same event ontology by identifying their output trigger overlap. Specifically, for an event ontology tree $\{\mathbf{e}_1, \mathbf{e}_2, ...\} \in \mathbf{E}$ with multiple event types and ten samples per event, we calculate the output trigger overlap ratio between two event types $\mathbf{e}_i, \mathbf{e}_j$ where $i \neq j$. The trigger overlap is measured by exact string matching each of the ten triggers in $\mathbf{e}_i$ with the ten triggers in $\mathbf{e}_j$. If the overlap ratio of output triggers exceeds a certain threshold (in our implementation, it is 0.5), we will consider one of the two events as duplicate and remove it from our dataset. This way, we can guarantee that the event types and output triggers of our dataset are diverse.

Generated Samples Table 13, Table 14 and Table 15 present examples of DivED dataset.

C Templates for Experiments

The compared baselines include ChatGPT, ChatIE (Wei et al., 2023) and (Gao et al., 2023a).

- **ChatGPT**: ChatGPT were prompted with the proposed method for a fair comparison with our finetuned LLaMA-2-7B. Prompt template is provided in Table 12.
- **ChatIE (Wei et al., 2023)**: ChatIE is a framework that transforms the zero-shot event detection task into a multi-turn question-answering problem. Here, LLMs are first prompted (as shown in Table 11)to identify the event type and then sequentially prompted to identify the trigger.
- **Gao et al. (2023a)**: This work explores the feasibility of ChatGPT as a zero-shot event detection model and further analyses the impact of event definitions, in-context examples and counterfactual examples in the prompt template in Table 10. We prompt ChatGPT with event definitions and positive examples in our implementation as this setup performed best on Gao et al. (2023a) evaluation.

Instruction:
Now I will give you some ontology, and then you need to output their definitions. Your output should carefully focus on the difference between these events.

In-Context Demonstration:
Input:
parent event: ...
son event: ...

Output:
parent event: event definition of parent event ...
son event: event definition of son event ...

Instruction:
Now I will give you some event ontology, and your output should carefully follow the previous examples. Your output should carefully focus on the event ontology.

Query:
Input:
parent event: mental state
son event: belief
son event: doubt
son event: awakening
son event: emotional state
son event: intention
son event: mental confusion
son event: understanding

GPT Output:
Mental State: A mental state refers to an individual's current psychological condition or state of mind. It encompasses various emotions, thoughts, and cognitive processes that a person may experience at any given time.
Belief: Belief is a mental state characterized by the acceptance or conviction that something is true or exists, often without concrete evidence. It involves having confidence in the validity of a particular idea, concept, or proposition.
Doubt: Doubt is a mental state characterized by uncertainty or skepticism regarding the truth or validity of a belief or idea. It involves questioning or being unsure about the accuracy of a particular belief or assertion.
Awakening: Awakening refers to a profound shift in an individual's perception or understanding. It is often associated with a realization or enlightenment, leading to a significant change in one's perspective or awareness, often of a spiritual or philosophical nature.
Emotional State: An emotional state relates to a person's feelings and emotions at a specific moment. These states can include happiness, sadness, anger, fear, or any other emotional response to internal or external stimuli.
Intention: Intention is a mental state involving a purposeful and deliberate desire to perform a particular action or achieve a specific goal. It is often accompanied by a plan or strategy to accomplish the intended outcome.
Mental Confusion: Mental confusion is a mental state characterized by a lack of clarity or cognitive disarray. It involves difficulty in processing information, making decisions, or understanding one's surroundings, leading to a sense of disorientation.
Understanding: Understanding is a mental state involving comprehension and insight into a concept, idea, or situation. It signifies the ability to grasp and make sense of information or knowledge, often leading to increased knowledge or awareness.

Table 7: Template for Ontology-Aware Event Definition Curation

Instruction:
Here we provide the definition and examples of three event types. The word between <trigger> and <trigger> is the trigger. The trigger must be a verb. Your output should be very diverse.

In-Context Demonstration:
Input:
Event type: ...
Event Definition: ...

Output:
Sample 1: ...
Sample 2: ...

Instruction:
Please generating a new event type following the same format. Note that there is an ontology, we provide event types and their definitions.

Instruction:
Please generate samples for event rebellion. Your output should carefully consider the ontology. Your output examples should not be examples for other son events.

Query:
Event Ontology:
parent event: historical event
son event: ceasefire
son event: cornonation
son event: rebellion
son event: revolt
Event Type: rebellion
Event Definition: Rebellion is an organized and sustained attempt to challenge an established power structure, often involving large numbers of people and various forms of resistance, including civil disobedience and armed struggle.

GPT Output:
Sample 1: The American Revolution was a <trigger>rebellion<trigger> against British colonial rule and led to the establishment of the United States.
Sample 2: The Syrian Civil War began as a <trigger>rebellion<trigger> against the government of President Bashar al-Assad.
Sample 3: The Boxer Rebellion was a failed <trigger>rebellion<trigger> against foreign influence in China in 1900.
Sample 4: The Haitian Revolution was a successful <trigger>rebellion<trigger> led by enslaved people against French colonial rule.
Sample 5: The Glorious Revolution of 1688 was a peaceful <trigger>rebellion<trigger> in England that resulted in the overthrow of King James II and the establishment of a constitutional monarchy.
Sample 6: The Arab Spring uprisings were a series of <trigger>rebellions<trigger> across the Middle East and North Africa in 2011 that aimed to overthrow authoritarian regimes.
Sample 7: The Boston Tea Party was an act of <trigger>rebellion<trigger> by American colonists in 1773 against British taxes on tea.
Sample 8: The Indian Rebellion of 1857 was a failed <trigger>rebellion<trigger> against British colonial rule in India.

Table 8: Template for Ontology-Aware Sample Curation

Instruction:
Now I will give you an event ontology and its corresponding event definitions for each event. You need to output more event definitions for the target event. Your output should carefully focus on the difference between these events in the event ontology. The output event definitions for the target event should be different from event definitions for other events. And you need to try to give me diverse event definitions for the target event, but maintain the original semantics.

In-Context Demonstration:
Input:
parent event: ...
son event: ...

Output:
1. ...

Instruction:
Now I will give you some event ontology, and your output should carefully follow the previous examples. Your output should carefully focus on the event ontology.

Query:
Input:
parent event: Mental State: A mental state refers to an individual's current psychological condition or state of mind. It encompasses various emotions, thoughts, and cognitive processes that a person may experience at any given time.
son event: Belief: Belief is a mental state characterized by the acceptance or conviction that something is true or exists, often without concrete evidence. It involves having confidence in the validity of a particular idea, concept, or proposition.
son event: Doubt: Doubt is a mental state characterized by uncertainty or skepticism regarding the truth or validity of a belief or idea. It involves questioning or being unsure about the accuracy of a particular belief or assertion.
son event: Awakening: Awakening refers to a profound shift in an individual's perception or understanding. It is often associated with a realization or enlightenment, leading to a significant change in one's perspective or awareness, often of a spiritual or philosophical nature.

Instruction:
Now please give me 10 more definitions for the event Awakening: Your output event definitions should be diverse, but maintain the semantics of the original definition. Your output should follow the below format.

GPT Output:
1. Awakening is a transformative experience marked by a deep and sudden insight that fundamentally alters a person's way of thinking or viewing the world, frequently triggering a newfound sense of purpose or clarity.
2. An Awakening signifies a profound shift in an individual's consciousness, usually triggered by a moment of profound understanding or revelation, which can lead to a complete reevaluation of their beliefs and values.
3. The concept of Awakening encompasses a powerful, introspective moment where an individual experiences a heightened awareness of their existence, often bringing about a renewed sense of meaning and direction in life.
...

Table 9: Template for ontology-aware event definition addition

Instruction:
This is an event extraction task where the goal is to extract structured events from the text. A structured event contains an event trigger word and an event type.

Event Types and their definitions:
Event type 1 : Event Definition 1
Event type 2 : Event Definition 2
:
:

In-Context Examples from the dataset
Positive Example 1
Sentence 1: ...
Output: Trigger, Event Type
Positive Example 2
Sentence 2: ...
Output: Trigger, Event Type
Positive Example 3
Sentence 3: ...
Output: Trigger, Event Type

Example 4
Sentence : User Query
Output:

Table 10: Template for Guo et al

Instruction:
The list of event types: <list all event types for the dataset>
Give a sentence: <user query>.
What types of events are included in this sentence? Please return the most likely answer according to the list of event types above. Require the answer in the form: Event type.

GPT Response:
Event Type

Instruction:
If the event type is identified, return the event trigger word(s). Return 'NONE' if the event type is absent. Separate multiple event trigger words with semicolon (;). Refrain from explaining your reasoning—provide only the direct answer.
Trigger -

Table 11: Template for multi-turn ChatIE

Instruction:
Act as an AI assistant specialized in extracting events. When given a sentence, a specified event type, and its definition, examine the sentence for the event type. If the event type is identified, return the event trigger(s). Return 'NONE' if the event type is absent. Separate multiple event triggers with semicolon (;). Refrain from explaining your reasoning—provide only the direct answer.

Sentence: <Sentence>
Event Type: <Type of Event>
Event Definition: <Definition of the Event>

TRIGGER:

Table 12: Template for GPT-3.5-Turbo. We prompt the model with the definition of each possible event type from the dataset and aggregate the predictions for evaluation.

Event Type:
ceasefire

Event Ontology:
Parent: historical_event
Sons: ceasefire, coronation, rebellion, revolt

Event Definition:
Definition 1: A ceasefire is a temporary cessation of armed conflict marked by an agreement between warring factions to lay down their weapons and cease all hostile activities, often in pursuit of a negotiated settlement or peace agreement.

Definition 2: Ceasefire is a legal agreement between two or more conflicting parties to temporarily halt hostilities, usually to allow humanitarian aid to reach the affected civilian population or to negotiate a long-term peace agreement.

Definition 3: A ceasefire refers to a state of truce or temporary peace between warring factions, allowing time for diplomatic negotiations and discussions to take place in pursuit of a more sustainable cessation of violence.
...
Definition 9: Ceasefire denotes a moment of respite in fighting between belligerent groups, often created through negotiations, that allows for the provision of humanitarian aid and the establishment of channels for peacebuilding and reconciliation.

Definition 10: Ceasefire is a crucial tool in conflict resolution and peacebuilding, serving as a vital step toward addressing underlying conflicts and arriving at a more permanent peace settlement.

Event Samples:
Sample 1: In 1991, the Persian Gulf War ended with a ceasefire.
Trigger: ceasefire

Sample 2: The two warring factions in the region agreed on a temporary ceasefire to allow humanitarian aid to reach the affected areas.
Trigger: ceasefire

Sample 3: After weeks of intense fighting, the UN brokered a ceasefire between the government and rebel forces.
Trigger: ceasefire
...
Sample 9: The military forces of two countries agreed to a ceasefire to allow for the exchange of prisoners of war.
Trigger: ceasefire

Sample 10: The two neighboring countries agreed to a ceasefire to de-escalate tensions and engage in peace talks.
Trigger: ceasefire

Table 13: Examples for the generated data for event ceasefire.

Event Type:
Change_event_time

Event Ontology:
Parent: Change_event_time
Sons: Holding_off_on, Change_event_duration

Event Definition:
Definition 1: A Change_event_duration is an event where the original duration of an activity or event is modified, either by increasing or decreasing the allotted time, to ensure the completion of the task or event.

Definition 2: Change_event_duration is an event that entails modifying the estimated duration of a particular activity or event based on assessment or evaluation data, such as delays, technical difficulties, or resource constraints.

Definition 3: Change_event_duration refers to the event of making revisions to the originally planned duration of an activity or event, typically done to accommodate changing priorities, shifting schedules, or other external factors.
...
Definition 9: A Change_event_duration is an event that involves adjusting the length of time allocated for a particular activity or event, motivated by a need to optimize efficiency, manage resources, or meet project objectives.

Definition 10: Change_event_duration refers to the event of extending or reducing the time frame for executing a particular task or activity, often done to accommodate shifting business needs or changing stakeholder demands.

Event Samples:
Sample 1: The concert promoters extended the length of the show due to popular demand.
Trigger: extended

Sample 2: The conference organizers shortened the duration of the keynote speeches to accommodate more panel discussions.
Trigger: shortened

Sample 3: The wedding planner adjusted the ceremony start time to avoid overlapping with the sunset.
Trigger: adjusted
...
Sample 9: The film festival prolonged its run for an extra day to showcase more entries.
Trigger: prolonged

Sample 10: The charity event shortened its fundraising campaign due to unexpected budget cuts.
Trigger: shortened

Table 14: Examples for the generated data for event Change_event_time.

Ontology: Parent: Arriving; Sons: Visiting scenario arrival, Drop in on, Access scenario
Parent: Arriving **Event Definition 1:** The act of Arriving involves the physical or virtual arrival at a destination or location, often involving anticipation and preparation for the event or activity that will follow. ... **Event Definition 10:** The Arrival event captures the moment when someone arrives at a particular location, often involving an emotional and physical shift as they transition into a new environment. **Sample 1:** The school field trip participants **arrived** at the museum and were greeted by the tour guide. ... **Sample 10:** The visitors **arrived** at the aquarium and were led to the dolphin show by the staff.
Son 1: Drop in on **Event Definition 1:** Drop in on refers to an unplanned and impromptu visit to a friend or acquaintance, often characterized by a surprise element and lack of formal invitations or arrangements. ... **Event Definition 10:** The act of drop in on signifies an unscheduled visit to an individual's place without prior notice or appointment, possibly to offer support or check on their well-being. **Sample 1:** Sarah decided to **pop in** on her friend who lived nearby and catch up. ... **Sample 10:** Jane had some free time on her hands and wanted to **pay a visit** to her former college roommate.
Son 2: Visiting scenario arrival **Event Definition 1:** Visiting scenario arrival entails arriving at a planned destination, such as a theater or concert, where specific events have been organized for the visitor's entertainment or education, creating a unique and memorable experience. ... **Event Definition 10:** The event of Visiting scenario arrival involves arriving at a location designated for a pre-planned gathering, such as a family reunion, where participants come together to socialize, network, or reconnect. **Sample 1:** The investors **arrived** at the company's headquarters for their business presentation. ... **Sample 10:** The family **reached** the theme park with their pre-booked ride tickets.

Table 15: Qualitative examples of DivED dataset. DivED contains diverse sibling events, high-quality samples with diverse triggers for each event type. The event definitions significantly distinguish the slight differences between sibling events.