FIG: Forward-Inverse Generation for Low-Resouce Domain-specific Event Detection

Anonymous ACL submission

Abstract

Event Detection (ED) is the task of identifying typed event mentions of interest from natural language text, which benefits domain-specific reasoning in biomedical, legal, and epidemiological domains. However, procuring supervised data for thousands of events for various domains is a laborious and expensive task. To this end, existing works have explored synthetic data generation via forward (generating labels for unlabeled sentences) and inverse (generating sentences from generated labels) generations. However, forward generation often produces noisy labels, while inverse generation struggles with domain drift and incomplete event annotations. To address these challenges, we introduce FIG, a hybrid approach that leverages inverse generation for high-quality data synthesis while anchoring it to domain-specific cues extracted via forward generation on unlabeled target data. FIG further enhances its synthetic data by adding missing annotations through forward generation-based refinement. Experimentation on three ED datasets from diverse domains reveals that FIG outperforms the best baseline, achieving average gains of 3.3% F1 and 5.4% F1 in the zero-shot and few-shot settings, respectively. Analyzing the generated trigger hit rate and human evaluation substantiates FIG's superior domain alignment and data quality compared to existing baselines.

1 Introduction

Event Detection (ED) (Sundheim, 1992; Doddington et al., 2004) involves identifying and categorizing significant events from natural language text based on a pre-defined ontology. It has applications in widespread domains like biomedicine (Pyysalo et al., 2012), epidemiology (Parekh et al., 2024b,c), law (Francesconi et al., 2010), etc. However, procuring annotations for each domain-specific ontology to train models is expensive and impractical. Although recent works introduce zero-shot LLM-reasoning (Gao et al., 2023; Cai et al., 2024), their performance falls short of supervised approaches (Huang et al., 2024).

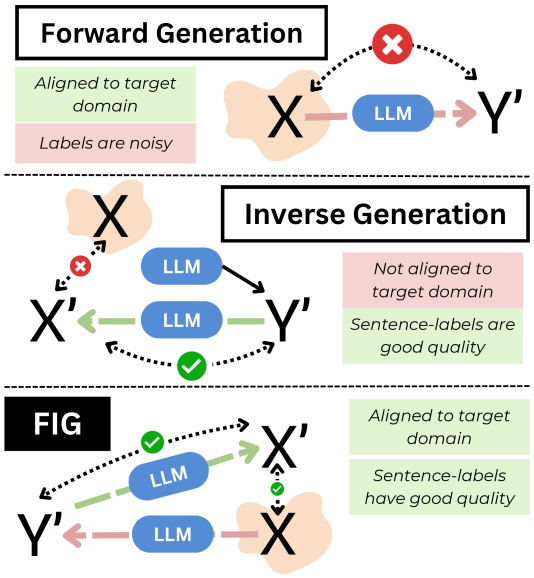


Figure 1: Comparing different data generation paradigms. Owing to poor LLM ED reasoning skills, forward generation (top) struggles with noisy labels. Inverse generation (middle) suffers from domain drift owing to lack of domain-specific cues. We introduce FIG (bottom), a hybrid forward and inverse approach, to overcome both issues to generate better quality data.

To mitigate the need for extensive annotations to train supervised models, prior work has explored LLM-powered synthetic data generation to enhance downstream model training. A widely used approach is a forward generation or weak supervision (He et al., 2021; Chia et al., 2022), where labels are assigned to existing unlabeled data (i.e., $X \rightarrow Y'$). This generation is limited by LLM ED reasoning, often leading to noisy labels. To address this, inverse generation has been proposed (Meng et al., 2022; Wang et al., 2023c), where sentences are generated based on sampled or generated labels (i.e $Y \rightarrow X'$). However, inverse generation introduces a domain drift between synthetic and target data

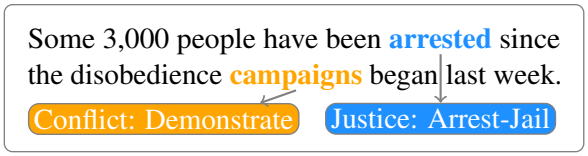
owing to the lack of any domain-specific cues, as illustrated in Figure 1.

In our work, we propose **F**orward-**I**nverse **G**eneration (FIG), a hybrid approach that integrates forward and inverse generation to enhance synthetic data quality. Instead of generating event triggers from scratch — an inherently high-variance process — we first apply forward generation to unlabeled data, extracting domain-specific triggers. Next, we utilize inverse generation to synthesize diverse sentences conditioned on the domain-specific triggers. Finally, a second forward generation annotates any missing events in the generated data to ensure high data quality. This approach produces cleaner synthetic data compared to forward generation while maintaining a closer domain alignment with target data than inverse generation, as shown in Figure 1.

We evaluate FIG on ED datasets from three different domains: ACE (Doddington et al., 2004) (news), SPEED (Parekh et al., 2024c) (social media), and GENIA2011 (Kim et al., 2011) (biomedical). Our primary method of evaluation is testing ED performance of DEGREE (Hsu et al., 2022) trained on the generated data. In the zero-shot setting, FIG outperforms STAR (Ma et al., 2024) (inverse generation baseline) and weak supervision (forward generation baseline) by an average of 16.8% and 3.5% Tri-C F1, respectively. Similarly, in the few-shot setting, FIG surpasses the strongest baseline by an average of 6.2% Tri-C F1. We study the generated trigger hit rate (relative to the gold trigger set) and demonstrate how FIG's triggers have about 7.3% better hit rate compared to the best baseline, which contributes to its superior performance. Finally, we also conduct human evaluation, which supports FIG's superior domain relevance and data annotation quality.

2 Problem Definition

We focus on the task of Event Detection (ED) (Sundheim, 1992; Doddington et al., 2004) for this work. The task of ED aims to extract mentions of any events of interest from natural language text. Following ACE 2005 (Doddington et al., 2004), we define an *event* as something that happens or describes a change of state and is labeled by a specific *event type*. The word/phrase that most distinctly highlights the occurrence of the event is defined as the *event trigger*, and the trigger-event type pair is referred to as the *event mention*. *Event Detection*, in particular, requires extracting the event *triggers* from the sentence and classifying it into one of the pre-defined event types. We provide an illustration of this task below where *arrested* and *campaigns* are the trigger words for the event types *Justice: Arrest-Jail* and *Conflict: Demonstrate* respectively.



In our work, we specifically focus on ED in diverse and specialized domains (e.g., biomedical), where procuring a training dataset D_T of annotated data points is expensive, whereas reasonable-sized unlabeled data D'_T is available. We focus on two realistic low-resource data setups - **zero-shot** (zero labeled data) and **few-shot** (k labeled datapoints per event type) settings. Different from domain transfer, we do not consider any labeled data for the source domain and directly optimize model performance for the target domain.

3 Methodology

In this work, we focus on Large Language Model (LLM)-powered synthetic data generation to alleviate the need for domain-specific annotated training data, as zero-shot LLM-based approaches usually struggle in ED in specialized domains (Huang et al., 2024). By generating a large amount of data instances $D_s = \{(X, Y)\}$, we can train downstream supervised ED models. Here, we first briefly review prior works on forward generation and backward generation. Next, we introduce our proposed data generation method FIG.

3.1 Background

We focus on two major paradigms for data generation in our work: forward generation and inverse generation. We describe them briefly below and provide illustrations in Figure 2.

Forward Generation: This is a more straightforward manner of data generation, wherein LLMs are utilized to generate labels Y' on unlabeled data $X \in D'_T$ (i.e. $X \rightarrow Y'$). This generation is analogous to *weak/distant supervision* (Weak Sup) (Mintz et al., 2009; Wang et al., 2021; He et al., 2021) where noisy/weak labels are assigned to clean sentences and then utilized to train down-

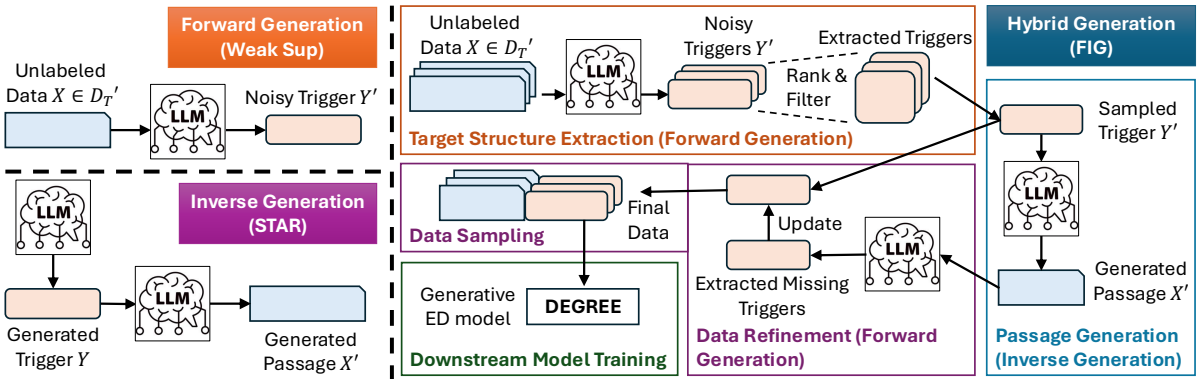


Figure 2: Model Architecture Diagram. Top left - illustration of forward generation. Bottom left - illustration of inverse generation. Right - FIG and its four components. We first extract domain-specific triggers via forward generation, then generate passages using inverse generation. Forward generation refines missing events, and we sample N data points per event for downstream training.

stream models. Consequently, the label quality is dependent on the reasoning capability of the LLM.

Inverse Generation: Inverse generation is a relatively new data generation paradigm (Kumar et al., 2020; Schick and Schütze, 2021). This paradigm first generates/samples potential labels Y based on the task definition and then generates plausible X' conditioned on the label Y (i.e. $Y \rightarrow X'$). Recent works like SynthIE (Josifoski et al., 2023) and STAR (Ma et al., 2024) have explored the usage of LLM-guided inverse generation for information extraction and event extraction tasks. Inverse generation provides control over data distribution via prompting, but the data quality is dependent on the LLM's sentence generation capability.

3.2 FIG

LLM reasoning has been poor for ED (Li et al., 2023a; Huang et al., 2024), which makes forward generation vulnerable to noisy label quality. Since LLMs are pre-trained to generate natural sentences, their generation capabilities are stronger, which supports inverse generation. However, owing to the lack of any domain-specific information, inverse generation can synthesize highly diverse sentences, leading to a domain drift. Merely specifying the domain in the prompt is not sufficient (shown in § D.1). Furthermore, inverse generation sentences can mention additional events which remain unannotated introducing noise in the data.

To this end, we propose **F**orward-**I**nverse **G**eneration (FIG), a hybrid approach that infuses forward generation reasoning into the base pipeline of inverse generation to ensure closer alignment to the target domain while improving the quality of the annotated labels. To procure domain-specific cues, we assume access to an unlabeled target domain data D_T' (similar to forward generation). We provide our architectural diagram in Figure 2 and explain each component of our pipeline below.

Target Structure Extraction: Similar to inverse generation, we first create potential triggers (i.e. labels) for each event type. However, generation from scratch can lead to very divergent set of triggers. For example, an "Attack" event can refer to a war in the news, cyberattacks in the cybersecurity domain, diseases in epidemiology, or criticism in economics. Naturally, each scenario would assume different triggers, which inverse generation struggles to generate even when provided with clear event definitions (actual examples in § 6.5).

Instead, we generate potential triggers per event type by utilizing an LLM to extract triggers using a forward generation over the unlabeled data D_T'. To ensure highly precise extraction of triggers, we develop a two-stage prompt setup. The first stage is tasked to identify and filter possible event types mentioned in the text based on the task instructions and event definitions. The second stage aims to find the most appropriate trigger word from the unlabeled sentence for each filtered event type. We illustrate this setup in § A.

After extraction, we aggregate and sort the triggers for each event type at the corpus level and filter out the top t triggers for each event type as the set of clean triggers. This filtering helps remove many noisy triggers introduced in the forward generation. Finally, we create target structures Y by sampling 1-2 event types per instance and sampling triggers from the clean trigger set for each sampled event

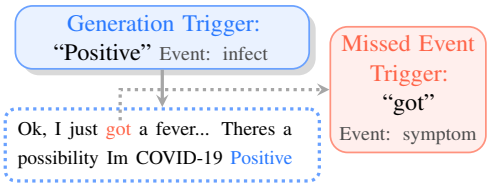


Figure 3: Illustration of how inverse generation can produce unannotated event mentions. Blue box = target event mention, red box = unannotated event mention.

type. We ensure uniform sampling of event types and triggers to avoid unbalanced label distribution. Overall, such forward generation-based trigger extraction ensures the distillation of domain-specific knowledge in the target structures.

Passage Generation: This component is the core of inverse generation and is tasked with generating passages corresponding to the constraints of the sampled target structure. Specifically, the LLM is prompted with the task instructions, the event definitions, the sampled target structure Y, and asked to generate a passage X' which mentions the event types using the triggers in Y. We illustrate this prompt in Figure 8. The key to generating better target data aligned passages lies in the presence of the target data aligned target structure Y (shown via examples in § 6.5).

Data Refinement and Sampling: While passage generation ensures the target structure events are mentioned in the passage, the passage could still have other unknown events which can lead to under-annotated (X', Y) data instances. We provide an illustration in Figure 3 with the target trigger *positive*, but the generated sentence also has a missing mention of *symptom* event triggered by *got*. To account for such missing annotations, we utilize forward generation to identify all possible event mentions in the generated passage. We remove duplicates and add the new event mentions to create a more complete target structure Y_f.

To further improve data quality, we apply an automated rule to remove passages that do not mention the target trigger. Additionally, we standardize trigger annotations by correcting variations in trigger word forms. Finally, we apply a greedy sampling algorithm to sample N instances (X', Y_f) per event type to create our final synthetic dataset D_s.

Downstream Model Training: The final component utilizes the generated synthetic data D_s to train downstream ED models in a supervised manner. The trained ED models are then used for inferring on the test set and for eventual evaluation.

4 Experimental Setup

Datasets: We consider three ED datasets from diverse domains for our experiments: (1) ACE (Doddington et al., 2004), in the news domain, (2) SPEED (Parekh et al., 2024c), in the social media domain, and (3) GENIA (Kim et al., 2011), in the biomedical domain. We simplify GENIA by converting the original document level annotations to sentence-level annotations. For the few-shot setting, we compile k data instances from the training data as the few-shot examplers.

For our unlabeled data, we consider two sources: (1) **Train** - annotation-free training splits (i.e. only the text) of each dataset and (2) **External** - unlabeled data from other external sources. For ACE, we utilize News Category Dataset (Misra, 2022) comprising Huffpost news articles from 2012-2022 as the external data source. We filter articles corresponding to political, financial, and business articles. For SPEED, we utilize COVIDKB (Zong et al., 2022) mining tweets from the Twitter COVID-19 Endpoint released in April 2020 as the external data source. Finally, we utilize GENIA2013 dataset (Kim et al., 2013) as the external data for GENIA. We provide statistics about these datasets in Table 8.

Baseline methods: We consider three LLM-based techniques for low-resource ED as the baselines for our work. (1) Inference (Gao et al., 2023): LLMs are used to directly infer on the target test data using their reasoning capability. (2) STAR (Ma et al., 2024): This model is the state-of-the-art inverse generation model for ED. It utilizes trigger generation, passage generation, and data refinement steps without using any unlabeled data, (3) Weak Supervision (Weak Sup): This model acts as the forward generation baseline. We utilize the Inference model to provide labels for the unlabeled data. For an upper bound reference, we also include a human generation baseline (Human) wherein we sample N data instances from the gold training data of each dataset to train the downstream ED model.

Base models: For our base LLMs, we consider three instruction-tuned LLMs of varying sizes, namely Llama3-8B-Instruct (8B model), Llama3-70B-Instruct (70B model) (Dubey et al., 2024),

Base LLM	Method	Unlabeled Data Source	ACE		SPEED		GENIA		Average	
			Eve-I	Tri-C	Eve-I	Tri-C	Eve-I	Tri-C	Eve-I	Tri-C
Llama3-8B	Inference	-	30.2	23.8	39.8	25.4	21.9	17.2	30.6	22.1
	STAR	-	44.9	35.0	21.0	10.1	25.9	19.0	30.6	21.4
	Weak Sup	train	41.7	37.8	45.6	31.5	26.9	21.4	38.1	30.2
	FIG(ours)	train	57.4	50.2	44.6	31.5	**35.2**	**28.9**	45.7	**36.9**
	FIG (ours)	external	**57.7**	**52.6**	**47.8**	**32.9**	33.6	24.6	**46.4**	36.7
Llama3-70B	Inference	-	46.9	41.3	46.9	35.6	34.2	28.2	42.7	35.0
	STAR	-	50.0	42.3	18.3	13.8	23.3	16.9	30.5	24.3
	Weak Sup	train	53.2	48.0	**52.8**	**39.6**	36.2	29.1	47.4	38.9
	FIG (ours)	train	58.1	53.8	49.9	38.7	38.0	29.7	48.7	40.7
	FIG (ours)	external	**59.7**	**55.6**	50.1	39.2	**39.2**	**31.5**	**49.7**	**42.1**
GPT-3.5	Inference	-	33.0	26.2	44.2	32.9	31.2	24.7	36.1	27.9
	STAR	-	45.0	36.6	21.3	14.6	21.8	14.3	29.4	21.8
	Weak Sup	train	49.7	44.6	**50.7**	**37.5**	37.7	30.1	46.1	37.4
	FIG (ours)	train	**54.8**	48.3	50.3	36.8	**39.3**	**31.1**	**48.1**	**38.7**
	FIG (ours)	external	54.0	**48.5**	50.1	36.1	38.7	29.4	47.6	38.0
-	Gold Data	-	64.6	61.6	64.0	53.5	51.3	44.0	60.0	53.0

Table 1: Zero-shot results comparing FIG with other baselines across three datasets and three base LLMs. Except for Inference, all other evaluations are performances of downstream DEGREE (Hsu et al., 2022) model trained on data generated by each technique. Eve-I: Event Identification F1, Tri-C: Trigger Classification F1.

and GPT-3.5 (175B model) (Brown et al., 2020). For our downstream ED model, we consider a specialized low-resource model DEGREE (Hsu et al., 2022), a generative model prompted to fill event templates powered on a BART-large pre-trained language model (Lewis et al., 2020).

Evaluation: Our primary evaluation metric is the synthesized data-trained model's ED performance on the final test splits of each dataset. We consider two low-resource settings - zero-shot (no labeled data) and few-shot (k datapoints per event type are used). Note this is different cross-dataset works (Cai et al., 2024) which train and test on an exclusive set of event types. For Inference, the LLM is directly run on the test set to procure model predictions. We report the F1 scores for two metrics (Ahn, 2006) for measuring model performance: (1) Event Identification (Eve-I) - correct identification of events, and (2) Trigger Classification (Tri-C) - correct identification of trigger-event pairs.

Implementation Details: We follow STAR for the implementation of the baseline models and the majority of hyperparameter settings. For FIG's passage generation, we select the top $t = 10$ triggers (except $t = 8$ for GENIA) for passage generation. We generate $N = 50$ datapoints per event type for each generation strategy. All our experimental results are reported over an average of three runs. Additional implementation details are provided in Appendix C.

5 Results

We present the results and insights for our zero-shot and few-shot settings below.

5.1 Zero-shot Results

We present the main results comparing the various methods across different LLMs and datasets in Table 1. For FIG, we show results using both the *train* data as well as the *external* data as the external data source. We highlight some of our main findings below.

Forward generation > Inverse generation: Although STAR performs decently for ACE, its performances for SPEED and GENIA are quite poor (even below the Inference baseline). Since SPEED (social media) and GENIA (biomedical) are more domain-specific than ACE (news), we attribute this poor performance to STAR's inability to generate domain-specific data instances (further validated in § 6.3). On the other hand, Weak Sup outperforms STAR, providing average gains of 13% Tri-C F1, thus establishing the superiority of forward generation-based synthetic data.

FIG performs the best: Our hybrid approach combines the benefits from both forward and inverse generation and provides the best performance. On average, FIG beats STAR by 17.3% Eve-I F1 and 16.3% Tri-C F1 points - proving how domain-specific cues from unlabeled data can help generate better-aligned inverse generated data. Compared to

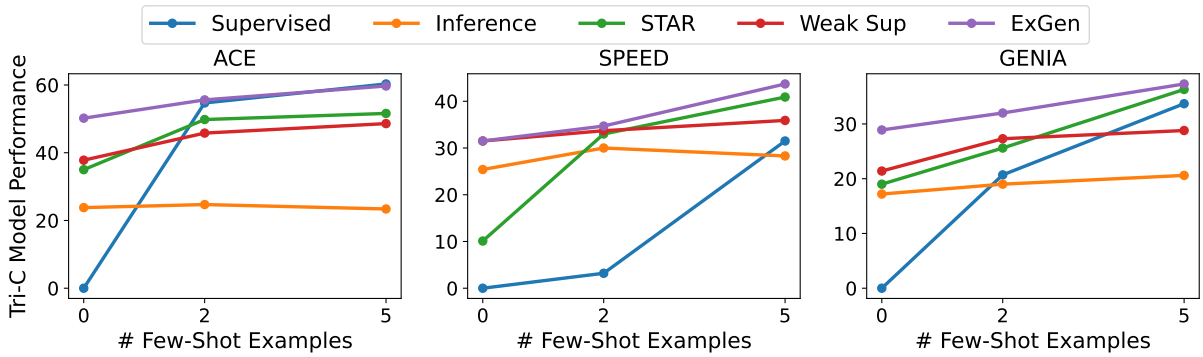


Figure 4: Few-shot results comparing FIG with other baselines across three datasets using Llama3-8B-Instruct as the base LLM. Except for Inference, all other evaluations are performances of downstream DEGREE (Hsu et al., 2022) model trained on data generated by each technique. Tri-C: Trigger Classification F1, #: Number of.

forward generation Weak Sup baseline, FIG provides average gains of 3.6% Eve-I F1 and 3.3% Tri-C F1 - suggesting how data diversity and cleaner label quality can help improve model performance.

External data source is effective: Assuming access to the training data as the unlabeled data source can be a strong assumption and bias for FIG. To cross-validate this assumption's impact, we also consider FIG with external data sources. Surprisingly, as seen from Table 1, FIG with external data provides similar gains of 4% Eve-I F1 and 3.4% Tri-C F1 over the best baseline. This demonstrates that instead of getting biased by external data, FIG utilizes the useful domain-specific signals (in the form of extracted triggers) from the external data and provides similar/better gains as with using the training data.

Generation ability doesn't scale up as reasoning ability: Although FIG performs better across all LLMs, the gains are much higher for Llama3-8B model relative to the Llama3-70B or GPT3.5 models. Comparing the improvements in model performance when scaling up from Llama3-8B to LLama3-70B, inverse generation-centric methods (STAR, FIG) improve by an average of 2.5% F1 points while reasoning-centric methods (Inference, Weak Sup) improve considerably better by an average of 10.8% F1 points. A similar disparity in performance improvements is observed for GPT3.5 as well. This sheds light on how the generation capabilities of LLMs do not scale compared to reasoning capabilities. Simultaneously, it shows that FIG is particularly effective when used with smaller LLMs whose the reasoning capability are poor.

5.2 Few-shot Results

We also study the various methods in the presence of small annotated data as part of our few-shot experiments. Specifically, we study the $k = 2$ and $k = 5$ few-shot settings, where k annotated examples per event type are utilized. Majorly, we utilize the k shots as in-context examples in the LLM prompts where applicable and append these few-shot examples to the synthesized training data as well. Additionally, we consider another baseline (Supervised) of downstream models trained only on the k shot examples. We present the Tri-C results for all the datasets for the Llama3-8B model in Figure 4 and summarize our major findings below.

FIG consistently performs the best: Similar to zero-shot results, we observe that FIG consistently beats all other baseline models. On an average, FIG outperforms STAR and Weak Sup by 5.4% Tri-C F1 and 7% Tri-C F1 respectively.

Generation abilities improve significantly: Contrary to our findings of small improvements when scaling up LLM model size in § 5, we observe that inverse generation-centric methods (STAR, FIG) consistently improve as we increase the number of shots. Contrastingly, reasoning-centric models stagnate and don't improve as much. This highlights how inverse generation becomes particularly effective in the presence of few exemplars.

6 Analysis

In this section, we provide various analyses to study FIG's superior performance. Unless specified, we utilize Llama3-8B-Instruct as the base LLM for the analyses.

Method	ACE	SPEED	GENIA
FIG	50.2	31.5	28.9
– Trigger Generation	43.2	27.8	28.2
– Data Refinement	47.4	23.3	22.0

Table 2: Ablation study for FIG's various components measured as Tri-C F1 performance across datasets.

Method	ACE		SPEED		GENIA	
	EI	TC	EI	TC	EI	TC
FIG	**57.4**	**50.2**	**44.6**	**31.5**	**35.2**	**28.9**
Weak Sup + STAR	46.9	38.9	44.5	29.5	30.2	24.3

Table 3: Comparing FIG's hybrid approach with data-mixing of forward and inverse generation.

Method	ACE	SPEED	GENIA
STAR	9.6%	15.3%	15.1%
Weak Sup	19.1%	38.4%	44.8%
FIG	**23.2%**	**49.4%**	**52.5%**

Table 4: Reporting the hit rate of synthetic data triggers relative to gold test triggers for the three methods.

Method	Naturalness	Event Relevance	Annotation Quality
STAR	3.1	3.4	3.1
Weak Sup	**4.2**	-	2.9
FIG	3.6	**4.0**	**3.6**

Table 5: Human evaluation for sentence naturalness, relevance of event in generated sentence, and the annotation quality. 1 = worst, 5 = best.

6.1 Ablation study

Table 2 shows the ablation study for each of our forward generation components. We observe how forward generation is critical in both trigger generation and data refinement stages with average performance reductions of 3.8% F1 and 6% F1 on removing the respective components from FIG.

6.2 Comparison with Data mixing

Data-mixing (Hoffmann et al., 2022; Xie et al., 2023) is a widely used technique for leveraging complementary information across datasets to promote robust downstream model training. We mix forward (Weak Sup) and inverse generation (STAR) as a hybrid baseline to compare with FIG. To keep comparisons fair, we consider $N/2$ data instances per event type from each dataset. Results from Table 3 demonstrate how FIG beats the data-mixing based hybrid model by 5-6% F1, highlighting the need for explicit model design to combine the benefits of forward and inverse generation.

6.3 Analyzing trigger quality and drift

ED models have a strong tendency to over-rely on lexical relations between triggers and events (Tong et al., 2022; Ma et al., 2024). Thus, we compare the synthetic data triggers with the gold test triggers as a raw study on the quality and drift of triggers in the synthesized data. Specifically, we extract triggers per event type in both datasets and measure the hit rate of the synthesized triggers on the gold set, as reported in Table 4. As observed, the poor hit rate shows the poor overlap of STAR's triggers with the gold triggers, which is a primary reason for its domain drift. Furthermore, the consistently stronger precision of FIG relative to Weak Sup demonstrate FIG's better trigger annotation quality.

To further study the label quality and relevance, we conduct a human evaluation. Specifically, a human expert ED annotator is tasked to score the generations (between 1-5) on the naturalness of the sentence for the domain, the correct relevance of the event mentioned in the generated sentence, and the annotation quality (more details in § D.3). We provide the averaged scores across the three datasets for 90 samples in Table 5. Weak Sup has high sentence quality but poor label annotations; while STAR suffers from poor event relevance indicating domain drift. Overall, FIG performs the best with high annotation quality and event relevance.

6.4 Domain-adapted LLM Fine-tuning

We fine-tune the base LLM used for passage generation on the unlabeled target-domain data D'_T to better align the generated passages to the target domain. Naturally, this can be applied only for smaller LLMs owing to fine-tuning costs. We present the results of fine-tuning Llama3-8B-Instruct on the unlabeled train data in Table 7. On average, we observe how target data fine-tuning improves FIG by a slight 0.5-2% F1, suggesting that target-domain passage generation may help, but it is not a highly influencing factor to improve downstream model performance.

6.5 Qualitative analysis of generated data

We provide qualitative evidence for FIG's domain-aligned triggers compared to STAR in Table 6 (more examples in Table 16). Owing to lack of domain grounding in STAR, the resulting triggers often appear misaligned (e.g. *asphyxiation* for *death*

Dataset	Event	Method	Trigger	Sentence
ACE	Attack	STAR	raid	As the rebels embarked on a daring trek across the desert, they launched a surprise **raid** on the heavily guarded fortress, catching the enemy off guard.
		FIG	shooting	As the rival businessman signed the contract, a sudden **shooting** erupted outside, causing chaos in the midst of the transaction.
SPEED	Death	STAR	asphyxiation	The hiker's life was tragically cut short as **asphyxiation** occurred after she became stuck in the narrow cave crevice.
		FIG	killed	The patient's feverish state was triggered when they tested positive for the virus, which ultimately led to their being **killed** by the rapidly spreading infection.
GENIA	Binding	STAR	merge	The regulatory protein's ability to activate a specific region of the DNA triggers the **merge** of two proteins, leading to the modification of gene expression.
		FIG	bound	During the phosphorylation of the enzyme, it **bound** to the DNA sequence, initiating the transcription process.

Table 6: Qualitative examples demonstrating STAR and FIG's trigger and sentence generation quality.

Method	ACE		SPEED		GENIA	
	EI	TC	EI	TC	EI	TC
FIG	**57.4**	50.2	44.6	31.5	35.2	28.9
+ SFT LLM	55.2	**51.7**	**46.9**	**35.8**	**36.7**	**29.1**

Table 7: Measuring model performance improvement using a LLM fine-tuned on unlabeled train data (SFT LLM) for FIG. EI: Event Identification F1, TC: Trigger Classification F1.

event related to pandemics). This misalignment carries over to the generated sentences, further reducing their quality and alignment. In contrast, FIG's triggers are better aligned to the target domain corpus, resulting in better quality data.

7 Related Works

Low-resource Event Detection Event Detection (ED) has been studied extensively (Sundheim, 1992; Grishman and Sundheim, 1996), leading to diverse datasets in news (Doddington et al., 2004; Song et al., 2015; Ellis et al., 2015), Wikipedia (Li et al., 2021; Pouran Ben Veyseh et al., 2022), and general domains (Wang et al., 2020; Parekh et al., 2023), as well as niche areas like biomedical (Pyysalo et al., 2012; Kim et al., 2011, 2013), multimedia (Li et al., 2020), cybersecurity (Satyapanich et al., 2020), epidemiology (Parekh et al., 2024b,c), and pharmacovigilance (Sun et al., 2022). To address the growing need for event detection across expanding domains, better low-resource domain-specific techniques are essential. Prior works have explored transfer learning via Abstract Meaning Representation (Huang et al., 2018), Semantic Role Labeling (Zhang et al., 2021), and Question Answering (Lyu et al., 2021). Reformulating ED as a conditional generation task has also aided low-resource training (Hsu et al., 2022, 2023; Huang et al., 2022). Recently, LLM-based reasoning (Li et al., 2023a; Gao et al., 2023; Wang et al., 2023b) and transfer-learning (Cai et al., 2024) has been explored for low-resource ED and transfer learning, but performance remains inferior to supervised models (Huang et al., 2024). This motivates efforts in synthetic data generation for low-resource ED.

Data Generation for Information Extraction LLM-powered synthetic data generation has been successful for various NLP tasks (Li et al., 2023b; Wang et al., 2023c; Wu et al., 2024; Shao et al., 2025). For information extraction, works have explored knowledge retrieval (Chen and Feng, 2023; Amalvy et al., 2023), translation (Parekh et al., 2024a; Le et al., 2024), and data re-editing (Lee et al., 2021; Hu et al., 2023). Some works in the directions we focus in our work include forward generation (Chia et al., 2022; Ye et al., 2022; Wang et al., 2023a; Tang et al., 2023) and inverse generation (Josifoski et al., 2023; Ma et al., 2024). We improve on existing forward and inverse generation works via our hybrid generation approach FIG.

8 Conclusion and Future Work

We introduce FIG, a hybrid forward-inverse generation approach for better domain-specific synthetic data in low-resource ED. Experiments on three diverse datasets reveal that forward generation suffers from noisy labels due to poor LLM reasoning, while inverse generation faces domain drift. FIG mitigates these issues, achieving superior domain alignment and cleaner data, leading to the best downstream ED performance in zero and few-shot settings. Future works can explore enhancing stronger domain alignment and extending FIG for multilingual generation.

Limitations

We consider only Event Detection (ED) as the main task for data generation, but our method can be extended to other structured prediction tasks as well. We leave this exploration for future works. We consider three specialized domains of news, social media, and biomedical to provide a proof-of-concept of our work. There are other specialized domains for ED as well which can be explored as part of future work. Finally, our proposed method FIGmakes a practical assumption of access to unlabeled data to procure target domain cues to guide the data generation. However, for specific super-specialized domains or if data has privacy concerns, this may not be possible and our method may not be applicable here. We assume such cases to be super rare and beyond the scope of our work.

Ethical Considerations

The theme of our work is to generate high-quality domain-specific data using Large Language Models (LLMs) using forward-inverse generations. The inherent LLMs can have certain biases, which can lead to potentially harmful or biased generations. Furthermore, the LLM can introduce potential hallucinations in the annotations which can hurt the model performance. We do not check or consider any bias/hallucination detection method as part of our work, as it is beyond the scope. So future users should take due consideration of this vulnerability.

Our proposed method FIG utilizes unlabeled data as a basis to procure domain-specific cues. If there are any biases in this data, it can propogate to the downstream model as well. We provide a proof-of-concept about our method in this work but do not detect or rectify such biases.

Since the inverse generation paradigm causes the LLM to generate sentences/passages, which can potentially be copied from the pre-training data the LLM has been trained on. This can potentially lead to copyright infringements and we do not consider any of such violations under consideration for our method. Users should consider this vulnerability before usage in commercial applications.

We would also like to mention and acknowledge that we have utilized AI chatbots to help with the writing of the work.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. Learning to rank context for named entity recognition using a synthetic dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382, Singapore. Association for Computational Linguistics.

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

Zefan Cai, Po-Nien Kung, Ashima Suvarna, Mingyu Ma, Hritik Bansal, Baobao Chang, P. Jeffrey Brantingham, Wei Wang, and Nanyun Peng. 2024. Improving event definition following for zero-shot event detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2842–2863, Bangkok, Thailand. Association for Computational Linguistics.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*.

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*, pages 45–57, Dublin, Ireland. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie

Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M. Strassel. 2015. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*. NIST.

Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors. 2010. *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Computer Science*. Springer.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2021. Generate, annotate, and learn: Generative models advance self-training and knowledge distillation. *CoRR*, abs/2106.06168.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and Philip S. Yu. 2023. Entity-to-text based data augmentation for various named entity recognition tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9072–9087, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4633–4646, Dublin, Ireland. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand. Association for Computational Linguistics.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. 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.

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, pages 1555–1574, Singapore. Association for Computational Linguistics.

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

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria. Association for Computational Linguistics.

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*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

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, Online. Association for Computational Linguistics.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. STAR: boosting low-resource information extraction by structure-to-text data generation with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18751–18759. AAAI Press.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *CoRR*, abs/2202.04538.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Rishabh Misra. 2022. News category dataset. *CoRR*, abs/2209.11429.

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

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024a. Contextual

label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Tanmay Parekh, Jeffrey Kwan, Jiarui Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, and Nanyun Peng. 2024b. SPEED++: A multilingual event extraction framework for epidemic prediction and preparedness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12936–12965, Miami, Florida, USA. Association for Computational Linguistics.

Tanmay Parekh, Anh Mac, Jiarui Yu, Yuxuan Dong, Syed Shahriar, Bonnie Liu, Eric Yang, Kuan-Hao Huang, Wei Wang, Nanyun Peng, and Kai-Wei Chang. 2024c. Event detection from social media for epidemic prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5758–5783, Mexico City, Mexico. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A novel multilingual event extraction dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Junichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinform.*, 28(18):575–581.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8749–8757. AAAI Press.

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*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Yichuan Ma, Peiji Li, Demin Song, Qinyuan Cheng, Shimin Li, Xiaonan Li, Pengyu Wang, Qipeng Guo, Hang Yan, Xipeng Qiu, Xuanjing Huang, and Dahua Lin. 2025. Case2Code: Scalable synthetic data for code generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11056–11069, Abu Dhabi, UAE. Association for Computational Linguistics.

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 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *CoRR*, abs/2303.04360.

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

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. 2023a. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147, Singapore. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Xingyao Wang, Sha Li, and Heng Ji. 2023b. Code4Struct: Code generation for few-shot event structure prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3640–3663, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *CoRR*, abs/2109.09193.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A survey on large language models for recommendation. *World Wide Web (WWW)*, 27(5):60.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*, pages 1331–1340, Online. Association for Computational Linguistics.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2022. Extracting a knowledge base of COVID-19 events from social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3810–3823, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Additional details about FIG

We provide illustrations about the various prompts used in FIG. For target structure extraction forward generation, we consider a two-prompt approach. The first prompt aims to identify if any events of interest are mentioned in the text as illustrated in Figure 6. It comprises the task definition, full event ontology with definitions, the task instructions, and the query sentence. The second prompt identifies the trigger corresponding to the event of interest, as illustrated in Figure 7. Here, we specify the task definition, the event ontology details, and the query text with the task instructions. To aid inverse generation for passage generation, we provide the task definition, the event ontology with event definitions, and the query comprising the sampled target structure. We illustrate this prompt in Figure 8. Finally, we provide the simplified one-prompt setup for forward generation utilized for data refinement in Figure 9.

B Additional Experimental Setup Details

B.1 Data Statistics

We discuss details about our dataset in § 4. Our test target domain data includes the test data splits of (1) ACE (Doddington et al., 2004) in the news domain, (2) SPEED (Parekh et al., 2024c) in the social media domain, and (3) GENIA (Kim et al., 2011) in the biomedical domain For unlabeled data, we utilize the training data of each dataset as one data source. For the other data source, we utilize data from external sources, specifically: (1) News Category Dataset (HuffPost) (Misra, 2022) comprising Huffpost news articles from 2012-2022 for ACE. We filter articles corresponding to political, financial, and business articles, (2) COVIDKB (Zong et al., 2022) mining tweets from the Twitter COVID-19 Endpoint released in April 2020 as the external data source, (3) GENIA2013 dataset (Kim et al., 2013) as the external data for GENIA. We provide statistics about this data in Table 8.

C Implementation Details

Here, we provide detailed implementation details for each component and models used in our work. We run most of our experiments on NVIDIA RTX A6000/A100 machines with support for 8 GPUs, while for GPT3.5, we make API calls through OpenAI.

Data Source	# Sents	# Event Mentions	Average Length
Test Data			
ACE - test	832	403	22.9
SPEED - test	586	672	28.1
GENIA - test	2,151	1,805	29.7
Unlabeled Train Data			
ACE - train	17,172	-	15.6
SPEED - train	1,601	-	33.5
GENIA - train	6,431	-	30.1
Unlabeled External Data			
HuffPost	43,350	-	17.4
COVIDKB	7,311	-	30.6
GENIA2013	6,542	-	17.4

Table 8: Data Statistics for the various test and unlabeled datasets used in our work.

C.1 LLM-based Generation

We provide details about the various hyperparameters for using LLMs in all the components of STAR and FIG. For Llama3-8B-Instruct and Llama3-70B-Instruct, we present the hyperparameters in Table 9; while Table 10 presents the hyperparameters for GPT3.5.

Batch Size	32
Temperature	0.6
Top-p	0.9
Max Generation Length	250

Table 9: Hyperparameters for decoding using Llama3-8B/70B model.

Base LLM	gpt-3.5-turbo-0125
Temperature	1.0
Top-p	1.0
Max Generation Length	500

Table 10: Hyperparameters for decoding using GPT3.5 model.

C.2 Few-shot Implementation Details

For the few-shot setting, we can access additional k datapoints per event type to aid better performance. For LLM-based prompting, we simply add these examples in the prompt as in-context examples to help the model do better reasoning/generation. For inverse generation methods (STAR, FIG), we do not add the k triggers to the extracted/generated trigger list, as it leads to a drop in model performance. This can be attributed to the presence of duplicate information as the trigger generation/extraction al-

ready accounts for the k triggers. For passage generation/Weak Sup, we append the k datapoints to the synthetically generated data to provide signals from the gold data.

C.3 Downstream Model Training

We choose DEGREE (Hsu et al., 2022) as our downstream model for evaluation, a generation-based prompting model that utilizes natural language templates. We implemented the DEGREE model under TextEE framework (Huang et al., 2024). Table 11 presents the primary hyperparameters for this model.

Pre-trained LM	BART-Large
Training Epochs	25
Warmup Epochs	5
Training Batch Size	32
Eval Batch Size	32
Learning Rate	0.00001
Weight Decay	0.00001
Gradient Clipping	5
Beam Size	1
Negative Samples	15
Max Sequence Length	250
Max Output Length	20

Table 11: Hyperparameters for DEGREE model.

C.4 LLM Fine-tuning

We discuss domain-adapted passage generation through LLM fine-tuning in § 6.4. Specifically, we conduct a low-rank finetuning (LoRA) (Hu et al., 2021) to reduce computational overhead to fine-tune Llama3-8B-Instruct. We implement LoRA using the `peft` and `trl` packages (Mangrulkar et al., 2022; von Werra et al., 2020). We choose the task of causal language modeling (i.e. continual pre-training) to perform domain adaptation on unlabeled in-domain sentences. We utilize cross-entropy loss on the dev split of the unlabeled data to select the best model. We provide additional details about the hyperparameters for this fine-tuning for each dataset in Table 12 below.

ACE	
Lora Rank	32
Lora Alpha	16
Lora Dropout	0.1
Learning Rate	0.0001
Weight Decay	0.05
Training Batch Size	32
Training Epochs	3
Eval Steps	20
SPEED	
Lora Rank	32
Lora Alpha	16
Lora Dropout	0.1
Learning Rate	0.00008
Weight Decay	0.05
Training Batch Size	32
Training Epochs	10
Eval Steps	20
GENIA	
Lora Rank	32
Lora Alpha	16
Lora Dropout	0.1
Learning Rate	0.00008
Weight Decay	0.05
Training Batch Size	32
Training Epochs	6
Eval Steps	20

Table 12: Hyperparameters for LoRA fine-tuning Llama3-8B-Instruct.

Method	**ACE**		**SPEED**		**GENIA**	
	EI	**TC**	**EI**	**TC**	**EI**	**TC**
STAR	**44.9**	**35.0**	**21.0**	10.1	25.9	19.0
+ mention	44.1	32.9	17.1	**10.3**	**28.7**	**20.4**
+ references	35.5	27.3	19.0	9.2	25.8	18.1

Table 13: Measuring model performance improvement providing domain-specific cues in the form of domain-mention (mention) or domain sentence references (references) to the LLM for STAR. EI: Event Identification F1, TC: Trigger Classification F1.

D Additional analyses

In this section, we provide additional analyses to support our main experiments.

D.1 STAR with domain-specific prompt

A simple way to infuse domain specific information in inverse generation pipelines like STAR would be to add domain-related information in the prompts to the LLM. We experiment with two such methods: (1) domain-mention, where we provide the target domain information in the prompt and ask the model to generate accordingly, and (2) domain-reference, where we use some examples from the unlabeled data in the prompt as reference sentences to better guide the passage generation. We provide results for these explorations using the Llama3-8B-Insturct model in Table 13. As observed, the results are generally poor, with an average drop of 0.1-0.6% F1 for domain-mention and 3.1-3.8% F1 for domain-reference. This is majorly because LLMs over-compensate, producing longer and more stereotypical information in their generations which hurts the naturalness of the sentence and causes further domain drift. Furthermore, the

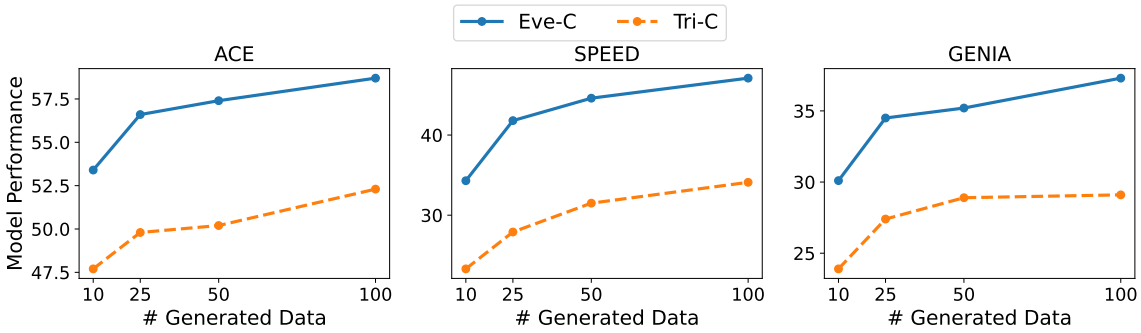


Figure 5: Model performance for FIG as keep change the number of generated datapoints N using Llama3-8B-Instruct for the three datasets.

Task Definition

You are an event detection system, looking to decide whether a sentence mentions or discusses an event from a specific list of events.

Full Event Ontology

Events of interests:
[infect, spread, symptom, prevent, control, cure, death]
An "infect" event is the process of a disease or pathogen invading host(s).
.....

Query

Does the following sentence discuss or mention any of the events of interest?

sentence: "If health officials tell people to wear masks to help stop the spread and save lives ....."

Figure 6: Prompt for stage 1 of forward generation for target trigger extraction.

Task Definition

You are an writer, looking to write sentences that contain specific events and event triggers. An event is a specific occurrence involving participants. An event is something that happens. An event can frequently be described as a change of state. Event trigger is the word that most clearly expresses its occurrence. Event triggers are often only a few words in length.

Related Event Ontology

An "infect" event is the process of a disease/pathogen invading host(s).
A "death" event signifies the end of life of individuals due to infectious disease.

Query

Generate a new sentence using trigger 'fumble' for event infect, trigger 'drowning' for event death.

Figure 7: Prompt for stage 2 of forward generation for target trigger extraction.

LLM makes more errors in mentioning the event as a part of its reasoning, which is utilized to make the generation in the domain style. We provide some qualitative examples for such generations in Table 14. In some ways, it also puts into light and amplifies the gains obtained by doing target domain SFT for FIG as discussed in § 6.4.

D.2 Impact of different number of training samples

We conduct a small analysis to study the impact of changing the number of generated samples on the downstream model performance for FIG. We present the results for Llama3-8B-Instruct in Figure 5. As observed, the performance continues to increase as we increase the data from $N = 10$ to $N = 100$ datapoints per event type. This promises that inverse generation will provide continued improvements by having larger control over data distribution.

D.3 Human Evaluation Details

We conduct a small human evaluation to judge the quality of the synthetic data in § 6.3. Here, we provide additional details about the human study and evaluation. Since the evaluation is conducted on three diverse and niche domains, we only utilize a single human annotator who is an ED expert and has previously worked on all three datasets as the primary annotator.

We majorly evaluate on three dimensions: (1) Sentence naturalness (SN): this metric judges

ACE
A 35-year-old cyclist was hit by a speeding car while riding to work, leaving her with severe injuries, while in a separate incident, a local retail giant filed a petition to restructure its debt, sparking concerns about its financial stability. As the war on terror raged on, the Mujahideen Advisory Council distributed a statement inviting Arab and foreign media reporters to enter Fallujah and cover the battles, while simultaneously, the ownership of the ancient artifacts was transferred to the museum, with the landlord demanding rent on the premises.
SPEED
As the influencer's viral challenge went viral, her followers were suddenly struck with a mysterious illness after the splash of a contaminated drink, leading to a shocking explosion of fatalities on social media. As the community struggled to come to terms with the devastating accident that had claimed the lives of several residents, the authorities swiftly implemented a strict quarantine to prevent the spread of the infectious disease, hoping to mitigate the tragedy.
GENIA
The specific transcription factor was elevated by the presence of the hormone, thereby increasing the expression of the target gene, while the inhibitory protein curbed the activity of a competing transcription factor, preventing the expression of a repressor gene. The binding of PEBP2/CBF to the promoter region boosts the expression of the gene, which turns on the production of a crucial cytokine in response to the immune response.

Table 14: Example passages of overly long and more stereotypical sentences generated when the domain is mentioned or references are added to the LLM prompt for STAR.

whether the sentence seems grammatical, natural, and fits the domain of the target data. (2) Event Relevance (ER): this metric is computed only for inverse/hybrid generation methods. This evaluation judges whether the sampled event and trigger are appropriately used to generate a sensible alignment with the target domain. Furthermore, it is verified if the right event definition is used. (3) Annotation Quality (AQ): this metric judges if the right trigger is used for each event mentioned in the synthetic output. If there are any missing events, then this score is penalized. For each metric, a score is given on a Likert scale (Likert, 1932) from 1 (worst) to 5 (best). We also provide event definitions for each event in each dataset as a reference for better judgment. We illustrate the annotation interface in Figure 10 and provide some sample examples in Table 15.

Task Definition

You are a writer, looking to extract a potential event trigger from a given sentence. Event trigger is the word that most clearly expresses the occurrence of the given event in the sentence. Event trigger is often only a single word in length.

Related Event Ontology

Event of interest: "spread"
A "spread" event is the process of a disease spreading/prevailing massively at a large scale.

Query

Given that the sentence mentions the event "spread", extract the trigger word in the sentence corresponding to this event type.

sentence: "If health officials tell people to wear masks to help stop the spread and save lives ....."

Figure 8: Prompt for inverse generation for passage generation.

D.4 Additional Qualitative Examples

In § 6.5, we discussed how FIG improves domain drift qualitatively relative to STARand provided some examples. Here, we provide more examples to further support that study in Table 16. This table further demonstrates how STAR can have a domain drift without a lack of domain-specific cues, while FIG is better here.

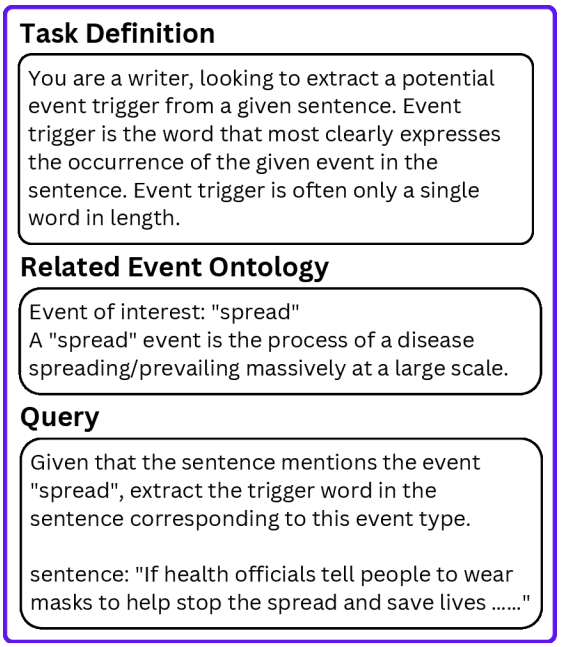


Figure 9: Prompt for forward generation for data refinement.

Sentence	Score
The sudden crash of the ambulance sent shockwaves through the hospital as medical staff rushed to the scene to monitor the patient's life signs, but it was too late, as the patient succumbed to the infectious disease.	SN: 2 ER: 1 AQ: 1
The wealthy entrepreneur transferred ownership of the struggling tech company to her trusted business partner, relinquishing control and financial responsibility	SN: 5 ER: 5 AQ: 5
Taken together, these data suggest that Id1 could be a possible target gene for mediating the effects of BMP-6 in human B cells, whereas Id2 and Id3 not seem to be involved.	SN: 4 ER: 3 AQ: 2

Table 15: Illustration examples for the human evaluation metrics. SN: sentence naturalness, ER: event relevance, AQ: annotation quality.

Rate all the metrics from 1-5. Use the filters on top to group by dataset and assign the scores
Naturalness = Is the snetence natural and grammatical?
Event Relevance = Based on event definitons (other sheet), figure if the event mentioned in this sentence seems correct
Annotation Quality = Check if all events are correctly annotated and there are no missing annotations.

Sentence	Annotation	Dataset	Naturalness of Sentence	Event Relevance	Annotation Quality
As the riot police stormed the square, they were met with an assault, and in the chaos, a protester's clothes caught fire, causing them to burn.	[{'event': 'Conflict:Attack', 'trigger': 'assault'}, {'event': 'Life:Injure', 'trigger': 'burn'}]	ACE			
The couple's marriage was annulled, ending their union after a tumultuous relationship.	[{'event': 'Life:Divorce', 'trigger': 'annulled'}]	ACE			
The court's decision was reconsider by the higher court after the losing party filed a petition to review the ruling.	[{'event': 'Justice:Appeal', 'trigger': 'reconsider'}]	ACE			
A dispute over a disputed contract led to a court proceeding being initiated, but the accused party was ultimately cleared of all charges.	[{'event': 'Justice:Sue', 'trigger': 'disputed'}, {'event': 'Justice:Sue', 'trigger': 'dispute'}, {'event': 'Justice:Acquit', 'trigger': 'cleared'}]	ACE			
The manager dial the CEO to discuss the quarterly sales report and provide an update on the marketing strategy.	[{'event': 'Contact:Phone-Write', 'trigger': 'dial'}]	ACE			
The entrepreneur's long-held vision was finally realized with the launch of her innovative startup, marking the beginning of a new chapter in her professional journey.	[{'event': 'Business:Start-Org', 'trigger': 'launch'}]	ACE			
As the company's financial struggles mounted, the CEO announced the exit of the firm, while the truck driver shifted gears and hit the road, hauling away the last remaining assets.	[{'event': 'Movement:Transport', 'trigger': 'shifted'}, {'event': 'Business:End-Org', 'trigger': 'exit'}]	ACE			
The accused was exonerate by the court, clearing their name of all charges, after a lengthy trial, and later, the two former colleagues met to chat and clear the air about their past misunderstanding.	[{'event': 'Justice:Acquit', 'trigger': 'exonerate'}, {'event': 'Contact:Meet', 'trigger': 'chat'}]	ACE			
The police officer arrested the suspect, initiating a court proceeding to determine the liability of the accused for the alleged crime.	[{'event': 'Justice:Sue', 'trigger': 'arrested'}]	ACE			
The candidate was chosen to lead the team after being selected by the majority of voters in the competitive election.	[{'event': 'Personnel:Elect', 'trigger': 'selected'}]	ACE			

Figure 10: Illustration of the interface for the human evaluation of the synthetically generated data. Short instructions are provided at the top. Each query comprises the sentence, annotation, and dataset. The human annotator is expected to score 1-5 for each of the three metrics on the right.

Dataset	Event	Method	Trigger	Sentence
ACE	Attack	STAR	raid	As the rebels embarked on a daring trek across the desert, they launched a surprise **raid** on the heavily guarded fortress, catching the enemy off guard.
			siege	The government issued fines to the rebel group, while the city was under **siege**, with gunfire erupting throughout the streets.
			slay	The wealthy businessman filed a complaint against his former partner, leading to a court proceeding, just as the rival gang leader **slays** his arch-nemesis in a brutal battle.
		FIG	shooting	As the rival businessman signed the contract, a sudden **shooting** erupted outside, causing chaos in the midst of the transaction.
			bombing	As the city was rocked by a devastating **bombing**, thousands of protesters took to the streets to demand justice and peace.
			fight	The rival gangs engaged in a brutal **fight**, sparking a wave of gunfire that left several people injured and buildings damaged.
SPEED	Death	STAR	asphyxiation	The hiker's life was tragically cut short as **asphyxiation** occurred after she became stuck in the narrow cave crevice.
			overdose	A young musician's **overdose** became the tragic finale of a life cut short in the midst of chaos.
			drowning	As she struggled to stay afloat, her heart stopped beating, and she succumbed to **drowning** in the icy waters.
		FIG	killed	The patient's feverish state was triggered when they tested positive for the virus, which ultimately led to their being **killed** by the rapidly spreading infection.
			lost	As the pandemic spread rapidly across the globe, thousands of people **lost** their lives due to the deadly virus.
			died	The elderly man, who had been suffering from a severe case of tuberculosis, **died** in his sleep.
GENIA	Binding	STAR	merge	The regulatory protein's ability to activate a specific region of the DNA molecule triggers the **merge** of two proteins, leading to the modification of gene expression.
			fuse	When the proteins **fuse** together, the activity of the transcription factor is inhibited, preventing the gene expression from proceeding.
			snap	When the two proteins **snap** together, the binding of the complex inhibits the expression of the target gene by deactivating a specific region of the DNA molecule.
		FIG	bound	During the phosphorylation of the enzyme, it **bound** to the DNA sequence, initiating the transcription process.
			translocation	The protein **translocation** to the nucleus triggers the induction of gene expression.
			binds	When the enzyme **binds** to the substrate, it activates the addition of a phosphate group to the target molecule, marking a crucial change in its function.

Table 16: Comparison of generated triggers and sentences from STAR and FIG methods