Is a Large Language Model a Good Annotator for Event Extraction

Appendix A: Statistical Analysis of the Benchmark Dataset

As highlighted in the paper, the current dataset faces significant challenges due to data scarcity and imbalance. Figure 1 presents the event data statistics for ACE 2005, while Figure 2 depicts those for MAVEN, which includes only the training and validation datasets since labels for the test dataset are unavailable. As evident from these figures, there's a pronounced long-tail issue in these datasets, with some event types lacking samples in both the development and test datasets.

Appendix B: Extended Experimental Analysis

In this section, we will provide a detailed examination of the EE strategies implemented on ACE 2005 and MAVEN using LLMs.

LLMs for joint EE on ACE 2005

As illustrated in Table 1, we evaluated the LLMs on joint EE, signifying simultaneous event detection and event argument extraction. As previously noted, the event detection performance drops significantly when considering both the prediction of event type and trigger word/phrase. Hence, in the joint EE, we focus solely on predicting the event type and its arguments. The prompt utilized for joint event extraction in EE is presented below:

LLMs	Tasks	Zero-shot Joint EE		
		P	R	F1
PaLM	ACE_ED	35.3	40.7	37.8
	ACE_EAE	19.7	21.1	20.4
GPT3.5	ACE_ED	52.8	40.2	45.6
	ACE_EAE	0.0	0.0	0.0

Table 1: Performance results for the zero-shot joint EE approach.

Please analyze the following sentence to determine if it contains any of the listed events: [...]. The specific role types of each event type are shown in this python dictionary ... If an event is detected, kindly provide its event type and corresponding arguments, formatting your response as: {Event_Type: event type} {Role_Type: argument text}. If no event is identified, simply return 'None'.

Sentence: A judge ruled in February that depositions in the divorce case will remain confidential .

Response:

Through our experiments, we discovered that PaLM accurately interprets this task, whereas GPT-3.5-Turbo doesn't produce any results related to event arguments. We further tested about 50 samples in GPT-4 and encountered the same problem. However, for solely event detection, GPT-3.5-Turbo outperforms PaLM. Interestingly, it identifies "previous marriages" as a trigger for the "Life.Divorce" event, but the correct label is "marriages" triggering "Life.Marry".

Utilizing LLMs for EE tailored to individual event types

Given the numerous event types in the benchmark dataset, including an example of each event type in the prompt might neither be efficient nor clear for LLMs. As such, we utilize PaLM to perform one-shot event extraction for each event type individually. The prompt applied in this scenario is presented below:

Sentence: You cite a session you had with President Lyndon Johnson in Austin , I guess , at his ranch , right ? Here is what you quote our President Johnson to you .

Event_type: Contact.Meet

Entity: Lyndon Johnson

Place: ranch

Given the above example, do you think the following sentence contains the event Contact.Meet? If no, please return 'None'. If yes, please show its specific arguments mentioned about this event. Format prediction as: {Event_type: Contact.Meet}; {Entity: entity}; {Place: place};. If no arguments mentioned, just output 'None' after each role type.

Sentence: Welch has previously failed in an attempt to temporarily seal a financial affidavit .

Prediction:

Initially, we assumed that this approach would outperform predicting all event types simultaneously. However, the event detection yielded an F1 score of only 21.6%, with the F1 score for event argument extraction being even

lower. Upon further investigation of the predictions, we discovered a high number of sentences identified as having the "Transaction.Transfer-Ownership" event. This anomaly arose because there were no training samples that covered all role types. As a result, we randomly selected an example and included it in the prompt. The example is presented below:

Sentence: BEGALA And how 'd they get them ?
Event_type: Transaction.Transfer-Ownership
Buyer: None
Seller: None
Beneficiary: None
Artifact: None
Place: None

This suggests that LLMs might perform better without an example than with a misleading one. It further underscores the importance of the prompt for LLMs.

LLMs for ED on MAVEN

We evaluated PaLM's event detection capabilities on MAVEN. The utilized prompt is presented below:

Please analyze the following sentence to determine if it contains any of the listed events: [[...]]. If an event is detected, kindly provide its event type and trigger word/phrase, formatting your response as: {Event_Type: event type} {Trigger: trigger word/phrase}. If no event is identified, simply return 'None'.

Sentence: The Swedish forces had for a long time laid siege to Stockholm , which was the last Danish stronghold in Sweden .

Response:

We observed that the majority of the predictions adhere to the required format. Nonetheless, the performance leaves room for improvement:

- Out of 9,400 test samples, PaLM failed to make predictions for 6 tests.
- It identified 3,182 event types that are not recognized within the MAVEN event spectrum, such as "Being_a_member" and "Opening".
- Upon submitting the predictions to the competition system[1], the performance metrics were as follows: precision at 21.8, recall at 6.9, and an F1 score of 10.5.

The unsatisfactory performance might be attributed to the overwhelming number of event types in the prompt. Evaluating the performance of each individual event type on popular LLMs should be a consideration for future work.

LLMs annotation performance

We utilized both PaLM and GPT-3.5-Turbo to generate additional training samples for ACE 2005. For event types with 300 or fewer labeled samples, we selected 5 from each. Our primary selection criterion was training samples that covered all roles. If no such samples existed or if they were insufficient, we chose 5 more training samples that, when combined, encompassed all roles. After discarding samples that didn't adhere to the required format, we obtained 148 labeled samples from PaLM and 172 from GPT-3.5-Turbo. The ED and EAE performances for different fine-tuning methods, with and without the enhanced dataset, are presented in Table 2 and Table 3. Notably, the majority of the enhanced data contributed to improved EE performance.

Method	Dataset	ED		
		P	R	F1
BERT+CRF	ACE	64.5	68.5	66.4
	ACE_GPT-3.5	64.5	74.4	69.1
	ACE_PaLM	67.3	74.2	70.6
DMBERT	ACE	61.6	75.2	67.7
	ACE_GPT-3.5	65.8	72.5	68.9
	ACE_PaLM	61.7	75.2	67.8
EEQA	ACE	63.8	74.9	68.9
	ACE_GPT-3.5	67.2	73.2	70.1
	ACE_PaLM	65.9	75.2	70.2
Text2Event	ACE	62.5	72.0	66.9
	ACE_GPT-3.5	61.5	71.0	65.9
	ACE_PaLM	63.2	71.7	67.2

Table 2: The comparison evaluates the performance of ED methods when fine-tuned with and without data augmentation. 'ACE_GPT3.5' represents the ACE 2005 training dataset augmented with labeled data from GPT-3.5. Conversely, 'ACE_PaLM' indicates the dataset enhanced using labeled data from PaLM.

Table 4 compares the performance of models trained solely on the MAVEN dataset with those trained on the enhanced dataset from LLMs. We ranked all event types based on the number of their training samples. For the event types ranked in the bottom 64, we randomly selected 10 training samples from each and utilized GPT-3.5-Turbo and PaLM to generate 10 new labeled samples. For event types ranked between 31 and 103, we randomly chose 5 training samples and enlisted LLMs to produce 5 new labeled samples for each. We implemented a filtering script to exclude samples that deviated from the required format or exhibited evident issues. While the table might not show significant performance improvements, we believe that with further investigation, the enhanced training samples could yield better results.

All finetuning experiments were conducted using the OmniEvent framework[2].

Appendix C: Challenges

Overall, EE presents a formidable challenge, particularly when working with limited and imbalanced labeled datasets. Figures 1 and 2 reveal a noteworthy trend: event types with fewer training samples tend to also have fewer test samples. This implies that the quantity and diversity of labeled samples, especially for specific events, play a crucial role in enhancing test performance—particularly for ACE 2005. Additionally, as highlighted by the authors of DYGIE++, the

[1] https://codalab.lisn.upsaclay.fr/competitions/395

[2] https://github.com/THU-KEG/OmniEvent

Methods	Triggers	P	R	F1
BERT+CRF	G_ACE	66.9	62.3	64.5
	P_ACE	50.1	64.9	56.5
	G_ACE_GPT-3.5	62.7	62.7	62.7
	P_ACE_GPT-3.5	46.1	63.4	53.4
	G_ACE_PaLM	67.4	65.6	66.5
	P_ACE_PaLM	49.5	66.7	56.8
DMBERT	G_ACE	64.1	71.5	67.6
	P_ACE	43.3	69.9	53.5
	G_ACE_GPT-3.5	65.4	71.5	68.3
	P_ACE_GPT-3.5	45.9	69.6	55.3
	G_ACE_PaLM	66.4	74.5	70.2
	P_ACE_PaLM	44.5	71.3	54.8
EEQA	G_ACE	70.7	54.2	61.4
	P_ACE	44.3	56.3	49.6
	G_ACE_GPT-3.5	71.4	58.9	64.5
	P_ACE_GPT-3.5	51.9	55.3	53.5
	G_ACE_PaLM	70.9	59.5	64.7
	P_ACE_PaLM	49.1	57.9	53.1
Text2Event	G_ACE	64.4	57.1	60.5
	P_ACE	44.2	55.1	49.0
	G_ACE_GPT-3.5	65.0	55.7	60.0
	P_ACE_GPT-3.5	45.5	55.2	49.9
	G_ACE_PaLM	64.0	54.3	58.8
	P_ACE_PaLM	45.8	54.1	49.6

Table 3: Performance evaluation for event argument extraction, considering data augmentation from both GPT-3.5 and PaLM. We distinguish between EAE outcomes based on golden triggers and those from prior ED predictions. To clarify, G_ACE denotes EAE performance using golden triggers from the original ACE 2005 dataset, while P_ACE_PaLM represents EAE results using predicted triggers from the ACE 2005 dataset augmented with PaLM.

Method	Dataset	ED		
		P	R	F1
SL	MAVEN	66.3	69.8	68.0
	MAVEN_PaLM	67.4	68.5	68.0
	MAVEN_GPT-3.5	69.9	66.3	68.0
Seq2Seq	MAVEN	63.3	64.4	63.9
	MAVEN_PaLM	62.8	65.0	63.9
	MAVEN_GPT-3.5	63.3	64.7	64.0

Table 4: The comparison evaluates the performance of the sequence labeling and sequence to sequence ED methods when fine-tuned with and without data augmentation on MAVEN.

small size and domain disparity between the development and test splits in the ACE 2005 dataset can render selections based on the development dataset unreliable. However, such issues might not be present in custom-built datasets or should be mitigated during their creation.

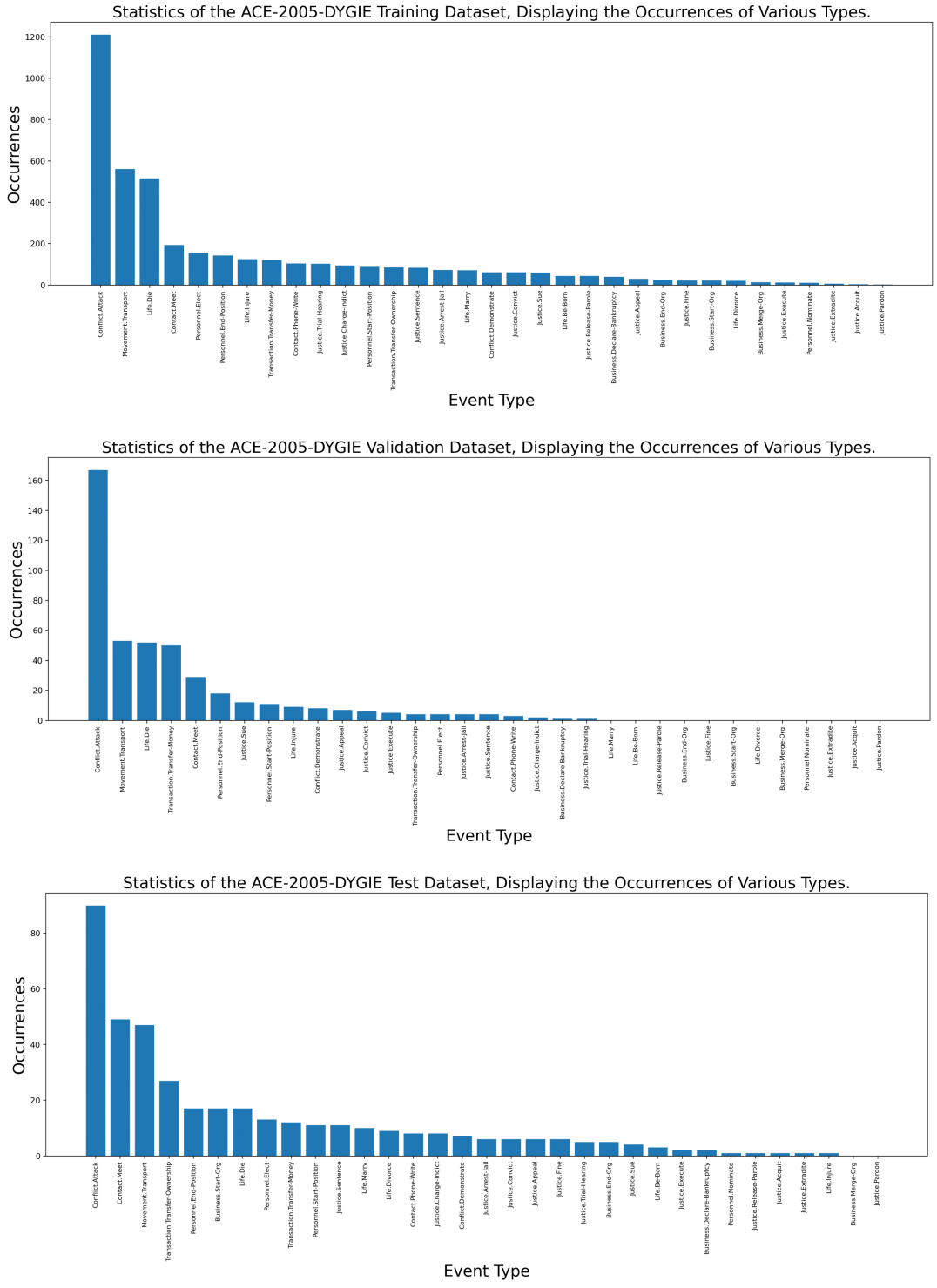

Figure 1: Event Type Occurrences in ACE 2005

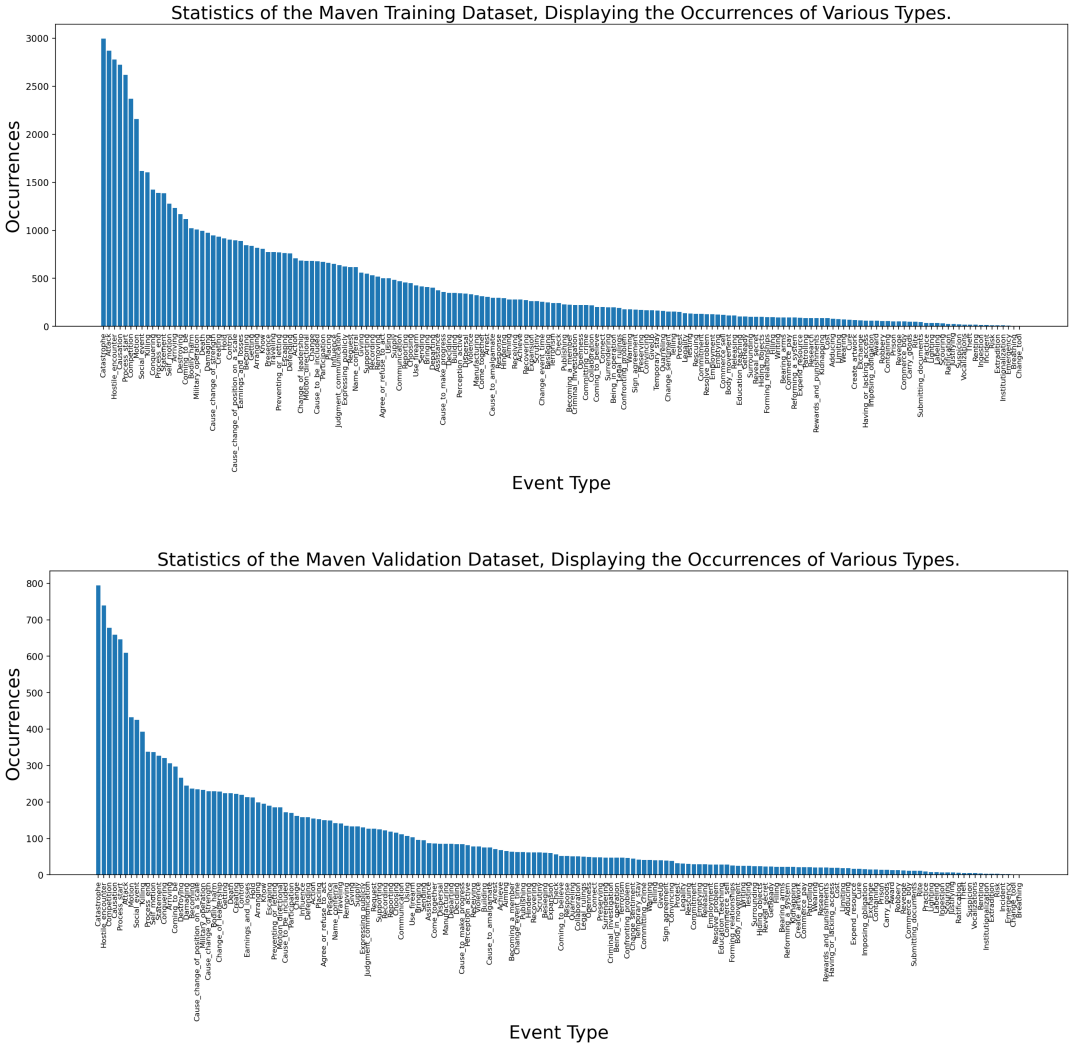

Figure 2: Event Type Occurrences in MAVEN