Are LLMs Good Annotators for Discourse-level Event Relation Extraction?

Kangda Wei, Aayush Gautam, Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University, College Station, TX
{kangda, aayushgautam, huangrh}@tamu.edu

Abstract

Large Language Models (LLMs) have demonstrated proficiency in a wide array of natural language processing tasks. However, its effectiveness over discourse-level event relation extraction (ERE) tasks remains unexplored. In this paper, we assess the effectiveness of LLMs in addressing discourse-level ERE tasks characterized by lengthy documents and intricate relations encompassing coreference, temporal, causal, and subevent types. Evaluation is conducted using an commercial model, GPT-3.5, and an open-source model, LLaMA-2. Our study reveals a notable underperformance of LLMs compared to the baseline established through supervised learning. Although Supervised Fine-Tuning (SFT) can improve LLMs performance, it does not scale well compared to the smaller supervised baseline model. Our quantitative and qualitative analysis shows that LLMs have several weaknesses when applied for extracting event relations, including a tendency to fabricate event mentions, and failures to capture transitivity rules among relations, detect long distance relations, or comprehend contexts with dense event mentions.

1 Introduction

Event Relation Extraction (ERE) refers to the NLP tasks that identify and classify relationships between events mentioned in a text. The commonly studied event relations include coreference, temporal, causal and subevent relations. ERE tasks aim to comprehend the intricate relationships between events and are beneficial for many applications, such as event prediction (Chaturvedi et al., 2017; Bai et al., 2021), question answering (Oh et al., 2017), and reading comprehension (Berant et al., 2014).

ERE tasks remain difficult and the empirical performance on these tasks are often rather low. Recently, inspired by the recent success of LLMs,

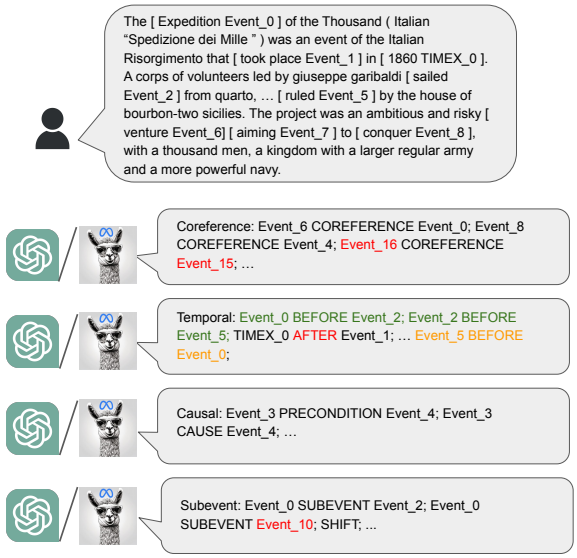


Figure 1: LLMs on MAVEN-ERE with four different relation types. MAVEN-ERE is an event relation extraction dataset covering four types of relations (coreference, temporal, causal, and subevent) with long documents. Pairs in orange indicate a violation of transitivity rules against pairs in green. Text in red indicates hallucinations.

Yuan et al. (2023) evaluates ChatGPT [1] on one ERE task, temporal relation extraction, using several benchmark datasets, including TDDiscourse (Naik et al., 2019) that addresses temporal relation extraction at the discourse level. However, Yuan et al. (2023) has to truncate the documents and constrain the size of discourse as inputs longer than eight sentences cause ChatGPT to generate unformatted answers. Meanwhile, Gao et al. (2023a) only evaluates ChatGPT on its causal reasoning ability. Unlike these works, we evaluate LLMs on performing multiple ERE tasks at the document level that tend to feature dense and long distance event relations, as shown in Figure 1. Specifically, we experiment with two LLMs, the commercial model GPT-3.5

[1] https://openai.com/blog/chatgpt

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1–19

and the open-source model LLaMA-2.

Furthermore, humans usually benefit from comprehending the meanings of event relations through examples (Hu et al., 2023) when conducting the ERE tasks, but the recent studies (Wei et al., 2023d; Yuan et al., 2023) often evaluate LLMs under the zero-shot setting and have a chance to limit the models in fully showcasing their true capabilities in event relation extraction. To fill the gap, we run experiments in the one-shot or few-shot settings, we also run Supervised Fine-Tuning (SFT) experiments aiming to further enhance LLMs on performing ERE tasks.

We systematically evaluate the effectiveness of LLMs on extracting four common types of event relations (coreference, temporal, causal, and subevent) using the `MAVEN-ERE` dataset (Wang et al., 2022) by experimenting with multiple prompts. The design of our prompts were informed by the prior work (Yuan et al., 2023; Bohnet et al., 2023) to ensure the most effective prompts applied. Comprehensive analysis, both quantitative and qualitative, lead to the following findings:

- **Performance:** Even with careful prompting, both GPT-3.5 and LLaMA-2 significantly underperform the baseline established through full supervised learning, and this is true for all the four types of relations. Although Supervised Fine-Tuning (SFT) can yield improved performance for LLaMA-2, there is still a gap between its performance and the performance of the baseline model trained with the same size of data, not to mention SFT for LLaMA-2 requires much more time and computation.

- **Transitivity and hallucinations:** There is a discernible tendency for LLMs to violate the rules of transitivity among event relation predictions. Furthermore, LLMs display inconsistencies in adhering to the provided prompts. Both suggest a potential lack of reasoning abilities and inaccurate understanding on the assigned task.

- **Events Distance and Density:** LLMs encounter challenges in capturing long distance event relations and inter-sentence event relations. LLMs also struggle to capture event relations in complex contexts that are dense with event mentions.

2 Related Works

Event Relation Extraction Event relation extraction (ERE) has been one of the fundamental challenges for natural language processing (Chaturvedi et al., 2017; Rashkin et al., 2018; Zhang et al., 2020). As understanding relations between events is crucial for understanding human languages (Levelt, 1989; Miller and Johnson-Laird, 1976) and beneficial for various applications (Khashabi et al., 2019; Zhang et al., 2020; Choubey and Huang, 2018), many approaches have been developed for performing ERE tasks (Liu et al., 2014; Hashimoto et al., 2014; Ning et al., 2017), and many high performing approaches are based on supervised learning (Dligach et al., 2017; Aldawsari and Finlayson, 2019; Liu et al., 2020; Lu and Ng, 2021).

LLMs for Extraction Tasks LLMs have been applied to several common information extraction tasks including event extraction, relation extraction and named entity recognition (González-Gallardo et al., 2023; Borji, 2023; Tang et al., 2023; Gao et al., 2023b; Wei et al., 2023d). However, to the best of our knowledge, LLMs have not been well explored for ERE tasks. Recently, Yuan et al. (2023) evaluates ChatGPT on temporal relation extraction and Gao et al. (2023a) evaluates ChatGPT on causal reasoning with the binary Event Causal Identification task, in contrast, we evaluate LLMs on extracting multiple types of fine-grained event relations. The dataset we use in this study, `MAVEN-ERE` (Wang et al., 2022), has dense relations at the discourse-level for four common types of event relations: coreference, temporal, causal, and subevent.

Prompt Engineering Many recent works have studied how to make LLMs perform better through applying various prompting techniques, including role-prompting (Zhang et al., 2023; Buren, 2023), re-sampling (Holtzman et al., 2020), one-shot or few-shot prompting (au2 et al., 2021; Shyr et al., 2023), and question decomposition (Wei et al., 2023c). Other novel and advanced techniques include Chain of Thought prompting (Wei et al., 2023a), least-to-most prompting (Zhou et al., 2023), learning from multiple prompts (Wei et al., 2023b), and retrieval augmentation (Lazaridou et al., 2022; Jiang et al., 2023). We refer to these techniques as guidelines when designing prompts in this work.

3 Experiment Setup

3.1 Data

For this study, we use the MAVEN-ERE dataset created by Wang et al. (2022), which includes annotations of four types of event relations: coreference, temporal, causal, and subevent. MAVEN-ERE consists of 4480 English Wikipedia documents, containing $103,193$ event coreference chains, $1,216,217$ temporal relations, $57,992$ causal relations, and $15,841$ subevent relations.

MAVEN-ERE is challenging as the documents contain comprehensive relation types, event relations at the discourse level and have denser relations among events comparing to other datasets. For example, MAVEN-ERE has an average of 272 temporal relation links per document comparing to 49 temporal relation links per document for MATRES (Ning et al., 2018); TimeBank-Dense (Cassidy et al., 2014) mainly focus on sentence-level relations; and TDDiscourse (Naik et al., 2019) only consider temporal relations.

As we test LLMs in a prompting setting without extra fine-tuning, we only utilize ten documents from the training set to extract examples included in a prompt, and we use ten documents from the validation set for prompt design and selection. We report the performance of LLMs on the whole test set of MAVEN-ERE , which contains 857 documents with $18,908$ event coreference chains, $234,844$ temporal relations, $11,978$ causal relations, and $3,822$ subevent relations.

3.2 Prompts

There are many possible ways to prompt LLMs for MAVEN-ERE . We design four different prompt patterns, namely Bulk Prediction, Iterative Prediction, Event Ranking referring to previous works (Yuan et al., 2023; Bohnet et al., 2023), and Pairwise. In the following sections, we describe each prompt pattern. Examples of prompt patterns can be found in Appendix D.

For all prompt patterns, we first label all event mentions as [x_i Event_p] where x_i is the triggering word in sentence S and Event_p is the Event ID given a document $D = \{S_1, S_2, ..., S_n\}$. TIMEX mentions are also considered for forming temporal relations, therefore, we label TIMEX mentions as [x_i TIMEX_q] and TIMEX_q is the TIMEX ID. p and q starts from 0 and gets increased by 1 each time a new event mention or TIMEX mention is labeled. We define E to be the set of event mentions and T to be the set of TIMEX mentions. We define Y to be the set of four relation types where $Y = \{$coreference, temporal, causal, subevent$\}$. R_y is defined to be the set containing all the sub-relation types for $y \in Y$, where $R_{coreference} = \{$COREFERENCE$\}$, $R_{temporal} = \{$BEFORE, CONTAINS, OVERLAP, BEGINS-ON, ENDS-ON, SIMULTANEOUS$\}$, $R_{causal} = \{$CAUSE, PRECONDITION$\}$, and $R_{subevent} = \{$SUBEVENT$\}$.

3.2.1 Bulk Prediction

Using the Bulk Prediction prompting, we query the LLM four times for each test document, with each query asking LLMs to list all relation pairs for each $y \in Y$. For each query, we also provide an example document followed by the gold relation pairs for the same y as the query. Notice this is a 1-shot setting since we provide a whole document as an example.

3.2.2 Iterative Prediction

Algorithm 1 sketches the Iterative Prediction prompting method. We query LLMs by iterating through the document D sentence by sentence. For each new sentence S, we append S to all the previous sentences that are already augmented with event relations predicted by the model. Each S is queried four times for each $y \in Y$. For coreference relation, we follow the Link-Append approach proposed by Bohnet et al. (2023) to augment the queried sentences. For temporal, causal, and subevent relations, we augment the sentences by inserting predicted relation tuples after the Event ID or TIMEX ID. A tuple is in the form $(e||t, r_y, e||t)$, where $e \in E$, $t \in T$, and $r_y \in R_y$.

We experiment with two ways for providing demonstrations to the model: (1) whole doc, and (2) n-shot.

Algorithm 1 Iterative Prediction

Inputs: Test Document D, Example Document $D^{'}$
for $y \in Y$ **do**
  for $S_i \in D$ **do**
    if $i == 0$ **then**
      Show an example with $D^{'}$;
      Query LLMs to list all relation tuples occurred in S_0 with relation $r_y \in R_y$;
      Augment S_0 with predicted tuples;
    else
      Show an example with $D^{'}$;
      Append S_i to augmented $S_{0:i}$
      Query LLMs to list all relation tuples occurred in $S_{0:i}$ with relation $r_y \in R_y$;
      Augment S_i with predicted tuples;

Whole Doc For each query, we provide one full training document augmented with gold event relations as an example. The length of the example increases as we iteratively go through the document sentence by sentence. Eventually, the model has access to the full example document augmented with gold relations for reference.

n-Shot For each query, we show n short documents augmented with gold relation labels as examples. Different from the Whole Doc approach described above, each document example here only consists of the first two sentences of an original training document. We retain the first two sentences of a document instead of only one sentence so that LLMs have access to event relations involving event pairs both within the same sentence and across different sentences.

Notice that the whole doc prompt utilize a whole document while the n-shot prompt only uses the first two sentences of a documents. We would like to provide whole documents for the n-shot prompt as demonstrations. However, the prompt length is constrained since there is a limitation for context length, which only allows one full document to be provided as demonstration. We designed the n-shot prompt since we would also like to explore the effects to LLMs by providing varying amounts of supervision.

3.2.3 Event Ranking

For the Event Ranking prompting method, we query LLMs by iterating through e and t for $\forall e \in E, \forall t \in T$ in test document D as shown in Algorithm 2. We ask LLMs to complete the query $(?, r_y, e||t), \forall e \in E, \forall t \in T, \forall r_y \in R_y$, and $\forall y \in Y$. Note that we only need to query TIMEX mentions for temporal relation since TIMEX mentions are only relevant to temporal relations. We also provide one example for each query. The example contains an example document, a query $(?, r_y, e^{'}||t^{'})$, where $e^{'}||t^{'}$ is an Event mention or a TIMEX mention from the example document, and the gold relation tuples as the answer. Notice the query for the test document, $(?, r_y, e||t)$, and the query in the example, $(?, r_y, e^{'}||t^{'})$, have the same r_y. This is also a 1-shot setting since we provide a whole document as an example.

3.2.4 Pairwise

For the pairwise prompting method, we query LLMs with all the event mentions and TIMEX mentions pairs. We ask LLMs to complete the query $(e||t, R_y =?, e||t), \forall e \in E, \forall t \in T$, and $\forall y \in Y$. Note that we only need to query TIMEX mentions for temporal relation since TIMEX mentions are only relevant to temporal relations. If there is no relation between two events then NONE should be predicted. We also provide one example for each sub-relation r_y of all four relation types. Note that this prompt pattern is in purely natural language format. However, since the number of event pairs equals the number of times we query LLMs, which grows quickly as the number of events increases, this approach is not feasible to use for GPT-3.5 and GPT-4 if we take into account of financial and time costs. Therefore, we only test this prompt with LLaMA-2.

Algorithm 2 Event Ranking

Inputs: Test Document D, Example Document $D^{'}$

for $y \in Y$ **do**
  if y == temporal **then**
    for $e \in E$ **do**
      for $r \in R_y$ **do**
        Show an example with D';
        Query LLMs with$(?, r, e)$ for D;
    for $t \in T$ **do**
      for $r \in R_y$ **do**
        Show an example with D';
        Query LLMs with$(?, r, t)$ for D;
  else
    for $e \in E$ **do**
      for $r \in R_{coreference}$ **do**
        Show an example with D';
        Query LLMs with$(?, r, e)$ for D;

3.3 Model

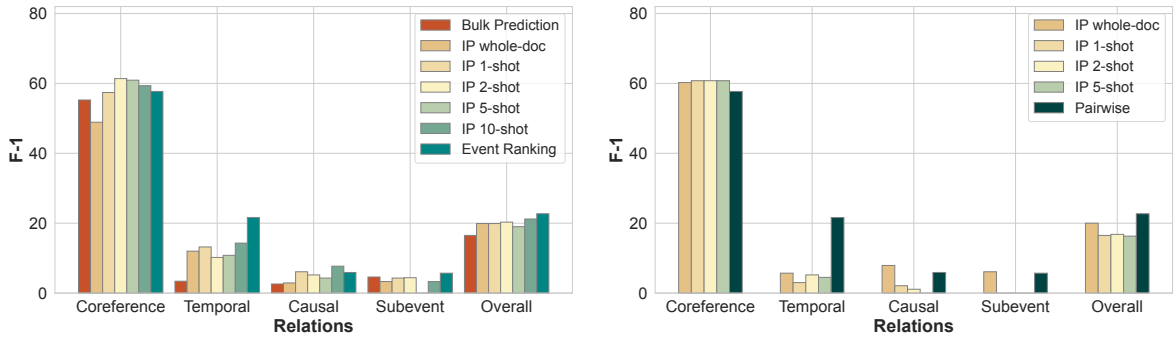
For this study, we use both open-source LLMs and closed-sourced LLMs for evaluation. For open-source models, we use *LLaMA-2 7B* model, specifically *Llama-2-7b-chat-hf*, accessed through Huggingface[2]. For closed-source models, we consider the *gpt-3.5-turbo-16k* model [3] from OpenAI API [4] as ChatGPT has been the most successful commercial LLMs so far. We test *Llama-2-7b-chat-hf* and *gpt-3.5-turbo-16k* on both validation and test sets using various different prompts.

To get an idea of how the newest GPT model performs, we also tested *gpt-4-1106-preview* model on a subset of the validation set instead of the whole test set becuase API calls to GPT-4 models are 30 times more expensive than GPT-3.5 models. The model performance was reported in Appendix E.

[2] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[3] More detailed information can be found in Appendix B

[4] https://platform.openai.com/docs/models



(a) GPT-3.5 (b) LLaMA-2

Figure 2: Comparisons of F-1 scores using different prompts on each type of event relations and the overall (macro average) performance, based on the results of GPT-3.5 (a) and LLaMA-2 (b) on the first 10 validation documents. The F-1 scores were calculated using the evaluation script provided by Wang et al. (2022), and the detailed results of precision and recall can be found in Appendix E.

Model	Prompt	hours	USD($)
GPT-3.5	Iterative Prediction	48	300
	Event Ranking	600	3,500
	Pairwise	3,600	30,000
LLaMA-2	Iterative Prediction	36	–
	Event Ranking	480	–
	Pairwise	3,000	–

Table 1: Estimated costs.

3.4 Prompt Decision

Before experimenting on the whole test set containing 857 documents, we test the four prompt patterns on the first 10 documents of the `MAVEN-ERE` validation set as running through the entire test set is both financially and time wise expensive. Table 1 shows our estimates of time and money needed to run each model through the whole test set using each of the latter three relatively costly prompts.

Figure 2 shows the results of GPT-3.5 and LLaMA-2 using different prompts on the first 10 validation documents[5]. The left sub-figure shows the results of GPT-3.5. We do not consider the Pairwise prompt for GPT-3.5 due to the extremely high cost on money. Among the remaining three prompts, Event Ranking achieves the best overall performance and Iterative Prediction performs relatively close. If we look into their performance over the relation types, Iterative Prediction wins on coference and causal relations while Event Ranking wins on the other two types of relations. Considering that Event Ranking is over ten times more expensive than Iterative Prediction (table 1), we choose Iterative Prediction over Event Ranking for running our full experiments.

The right sub-figure of Figure 2 shows the results of LLaMA-2. LLaMA-2 failed to generate consistent outputs when using Bulk Prediction and Event Ranking, thus we do not include those numbers. Between the Pairwise prompt and Iterative Prediction[6], the Pairwise prompt yields a slightly better overall performance. However, across the four relation types, the Iterative Prediction prompt, IP whole-doc, wins on three relation types (coreference, causal and subevent relations) while the Pairwise prompt only wins on temporal relations. In addition to performance, we also consider the dramatic running time differences between the two prompts (table 1), and we choose Iterative Prediction over the Pairwise prompt for running full experiments using LLaMA-2.

To summarize, we choose Iterative Prediction with its variations as the final prompts for both models when running the full experiments.

3.5 Supervised Baseline

We consider the baseline model proposed by Wang et al. (2022) as our baseline model. The baseline model utilizes *roberta-base* model from Huggingface[7] as the underlying language model and trains separate classification heads for each relation type $y \in Y$. The baseline model is trained end-to-end and performs pair-wise classification. We train the baseline model for each relation type separately for fair comparison against GPT-3.5, and LLaMA-2. We strictly follow the training and evaluating

[5]The results of GPT-4 on the first 10 validation documents are included in Appendix E.

[6]IP 10-shot was not performed for LLaMA-2 due to context length constraint.

[7]`https://huggingface.co/roberta-base`

processes according to Wang et al. (2022).

4 Results

LLMs Performance We report the performance of GPT-3.5 and LLaMA-2 on extracting coreference, temporal, subevent, and causal relations in Table 2. We can see that both models severely underperform the supervised baseline model in extracting each of the four types of relations.

For coreference relations, we report results using four metrics following previous work (Wang et al., 2022). Table 2 shows that among the IP prompt variants, the whole doc prompt yields the best coreference resolution performance for both GPT-3.5 and LLaMA-2. For the n-shot prompts, the performance decreases as n increases. It appears that by providing LLMs an example document with all the mentions and gold cluster included, the whole doc prompt hints LLMs to search through a whole document and better link coreferential mentions compared to the n-shot prompts that only provide excerpts from different documents.

Regarding temporal, causal, and subevent relations, GPT-3.5's performance increases as the number of example documents increases for n-shot prompt with the 10-shot prompt yielding the best performance. While for LLaMa-2, the whole doc prompt yields the best performance.

Regarding individual types of relations, both models have the best performance on temporal relations and the worst performance on subevent relations. We believe such a performance gap is primarily because temporal relations are much denser than subevent relations in `MAVEN-ERE` .

The last column of Table 2 shows the macro-average performance over the four types of relations. We can see that the 10-shot prompt performs slightly better than the whole doc prompt for GPT-3.5 while the whole doc prompt achieves the best performance for LLaMA-2.

SFT with a Varying Size of Training Data Will LLMs perform better when we fine-tune them with training data? How will LLMs compare to the smaller baseline model when fine-tuned using the same amount of data? We answer this question by varying the amount of training data for the supervised baseline method and meanwhile fine-tuning LLaMA-2with the same amount of training data. In this experiment, LLaMA-2 is fine-tuned in the pairwise format as described in Section 3.2.4 to match with the baseline model (Wang et al., 2022) which treats the ERE tasks as pairwise classification tasks. Detailed hyperparameters for fine-tuning LLaMA-2 can be found in Appendix A.

Figure 3 shows the result comparisons between LLaMA-2 and the smaller baseline model. First, SFT certainly improves the performance of LLaMA-2. However, it still underperforms the smaller baseline method for all the relation types as the number of used training documents increases. LLaMA-2 typically requires twice more training data to reach the same overall performance as the smaller baseline model. LLaMA-2 only outperforms the smaller baseline method when the available training data is very limited, as LLaMA-2 benefits from its zero-shot capability emergent from large-scale pre-training. It also deserves mentioning that fine-tuning LLMs is much more expensive than fine-tuning the smaller baseline language model *RoBERTa*. For example, fine-tuning using 200 training documents for 3 epochs requires approximately 72 hours for LLaMA-2 7B, while only about one hour is needed for the *roberta-base* baseline model.

5 Discussion

5.1 GPT-3.5 and LLaMA-2 Struggle to Follow the Prompt Consistently

During the test of GPT-3.5 and LLaMA-2, we notice that both models create events or event relations that do not exist in text.

In addition, both models occasionally have difficulties in generating formatted answers, and the outputs may consist of random words from the document rather than Event or TIMEX ID or even event trigger words.

As GPT-3.5 has achieved overall better performance compared to LLaMA-2, we conduct further analysis on model performance mainly based on the predictions of GPT-3.5 on the 10 validation documents.

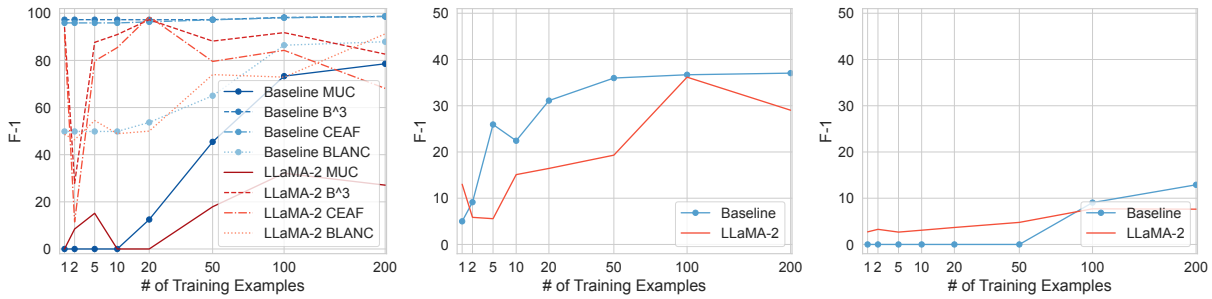
5.2 GPT-3.5 failed to learn transitivity rules

By manually examine the output of GPT-3.5 on the 10 validation documents, we notice that this model failed to learn the transitivity rules from the provided examples. When the output from GPT-3.5 contains tuples like (Event_0, BEFORE, Event_1) and (Event_1, BEFORE, Event_2), the tuple (Event_0, BEFORE, Event_2) can be inferred from the existing predictions. However, GPT-3.5 failed to include such tuples that can be inferred

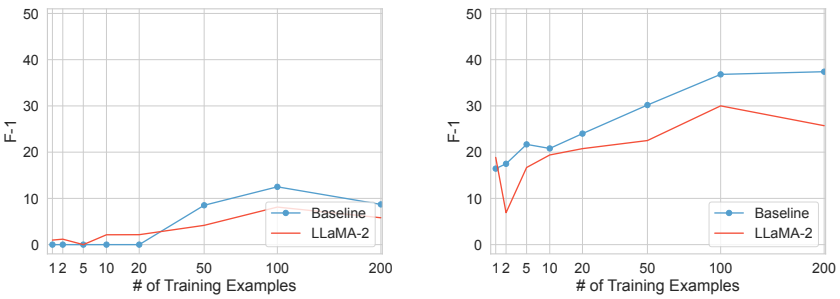
Iterative Prediction	MUC (Vilain et al., 1995) Precision	Recall	F-1	B^3 (Bagga and Baldwin, 1998) Precision	Recall	F-1	$CEAF_e$ (Luo, 2005) Precision	Recall	F-1	BLANC (Recasens and Hovy, 2011) Precision	Recall	F-1
GPT-3.5												
whole doc	**21.6**	**25.7**	**23.2**	91.7	**93.2**	92.5	91.6	89.9	90.1	**57.8**	**56.3**	**56.9**
1-shot	15.3	17.0	16.1	92.0	92.8	92.4	91.0	90.1	90.6	54.3	53.8	54.0
2-shot	17.9	18.9	18.4	92.6	92.9	92.7	91.2	90.7	91.0	54.9	54.3	54.5
5-shot	17.7	15.2	16.4	**93.9**	92.7	**93.3**	**91.9**	**92.3**	**91.7**	55.4	53.3	54.0
10-shot	11.5	12.0	11.8	92.3	92.5	92.4	90.5	90.2	90.4	53.2	52.2	52.6
LLaMA-2												
whole doc	**10.6**	**6.9**	**8.4**	95.1	**92.2**	93.7	**90.6**	93.4	92.0	**53.0**	**51.1**	**51.5**
1-shot	0	0	0	**100.0**	92.0	**95.8**	90.5	**98.4**	**94.3**	49.3	50.0	49.7
2-shot	0	0	0	100.0	92.0	95.8	90.5	98.4	94.3	49.3	50.0	49.7
5-shot	0	0	0	100.0	92.0	95.8	90.5	98.4	94.3	49.3	50.0	49.7
Baseline	$79.8_{1.6}$	$83.6_{0.5}$	$81.7_{0.7}$	$97.8_{0.2}$	$98.4_{0.0}$	$98.1_{0.1}$	$98.0_{0.1}$	$97.6_{0.2}$	$97.8_{0.1}$	$87.6_{1.1}$	$92.1_{0.1}$	$89.7_{0.6}$

Iterative Prediction	Temporal Precision	Recall	F-1	Causal Precision	Recall	F-1	Subevent Precision	Recall	F-1	Overall F-1
GPT-3.5										
whole doc	19.8	4.4	7.2	2.9	2.9	2.8	**1.9**	1.3	1.6	19.3
1-shot	17.1	4.5	7.1	4.1	2.7	3.3	1.9	1.2	1.5	18.8
2-shot	19.5	4.3	7.1	4.1	2.6	3.2	1.5	0.9	1.2	18.9
5-shot	21.3	5.8	9.1	4.5	3.0	3.6	1.7	1.4	1.6	19.5
10-shot	**26.8**	**8.0**	**12.3**	**5.3**	**5.3**	**5.3**	1.7	**2.8**	**2.1**	**20.4**
LLaMA-2										
whole doc	17.2	**3.1**	**5.2**	4.1	**5.0**	**4.5**	3.4	**6.3**	**4.4**	**18.9**
1-shot	15.4	1.2	2.2	4.6	0.2	0.3	3.3	0.1	0.2	15.7
2-shot	**26.3**	2.2	4.1	4.6	0.1	0.2	0	0	0	16.1
5-shot	19.4	1.3	2.4	**8.2**	0.2	0.3	**4.5**	0.2	0.4	15.8
Baseline	$57.3_{0.6}$	$54.5_{0.1}$	$55.8_{0.2}$	$34.2_{0.1}$	$29.3_{1.0}$	$31.6_{0.6}$	$29.5_{2.5}$	$25.4_{2.6}$	$27.2_{0.9}$	51.6

Table 2: Event coreference,temporal, causal, and subevent relations performances of GPT-3.5 and LLaMA-2 on MAVEN-ERE test set using Iterative prediction prompt patterns comparing to the baseline. The numbers in **bold** indicate the best performance across different prompt patterns for GPT-3.5 or LLaMA-2 separately. We report averages and standard deviations over 3 random seeds for the baseline method.



(a) Coreference **(b)** Temporal **(c)** Causal



(d) Subevent **(e)** Overall

Figure 3: F-1 scores for coreference, temporal, causal, subevent relations, and overall performance. The blue lines represent the baseline's performance trained with different numbers of documents. The red lines represent the performance of LLaMA-2 fine-tuned with different numbers of documents.

		FP	FN	Transitivity Fixable
whole doc	Temporal	21.63	66.33	7.40
	Causal	64.33	32.48	0
	Subevent	77.35	20.99	–
10-shot	Temporal	23.85	63.42	6.02
	Causal	59.71	34.29	0.43
	Subevent	82.77	15.55	–

Table 3: Rates (%) of various errors in GPT-3.5 predictions with Whole doc and 10-shot prompt settings. False positives (FP) and false negatives (FN) are indicated. Transitivity fixable denotes the enhancement in F-1 score achieved by incorporating transitive relations through true positives.

using transitivity rules. On the contrary, GPT-3.5 sometimes predict the opposite of the correct tuple. In this case, instead of (Event_0, BEFORE, Event_2), (Event_2, BEFORE, Event_0) will be predicted by GPT model, which clearly violates transitivity rules.

According to Wang et al. (2022), 88.8% temporal relations and 23.9% causal relations can be inferred with transitivity rules in the `MAVEN-ERE` dataset. Failure to follow the transitivity rules detriments GPT's performance. Moreover, Table 3 highlights that false negatives and false positives emerge as the predominant error types, indicating that GPT-3.5 faces challenges in accurately discerning the presence or absence of relationships. Notably, in the context of temporal relationships, a noteworthy increase in the F1 score is observed when incorporating all transitivity-inferred relations. This implies that GPT-3.5 indeed falls short of capturing a comprehensive array of transitive relations.

5.3 Event Pairs with a Varying Distance

	Intra (< 1)	Inter (>= 1)	1	2	3	4	>= 5
Temporal	25.45	12.39	19.97	17.70	10.23	11.76	6.87
Causal	8.92	7.02	11.02	7.69	0	0	8.70
Subevent	4.65	2.56	5.26	0	0	0	0

Table 4: GPT-3.5 performance (F-1 score) using Iterative Prediction 10-shot prompt on data groups with varying distances (measured in #sentences) between related events.

We investigate model performance on predicting intra- and inter-sentence event relations separately. Given that event coreference resolution relies on undividable clusters, we mainly analyze performance on the other three tasks. As shown in Table 4, GPT-3.5 is more capable to capture the relations between events that appear in the same sentence (intra-sentence) and otherwise struggles with capturing the inter-sentence event relations.

We also investigate the impact of sentence distance on model performance for inter-sentence cases. As shown in 4, the performance on temporal relation extraction decreases quickly as the number of sentences between two events increases; the performance on causal relation extraction also decreases a little when the number of sentences in between increases from one to two, but then the performance seems to remain stable when we further consider causal relations with five or more sentences in between. Overall, we observe lower performance on event pairs with greater distances.

5.4 Event Pairs in Contexts of Varying Event Densities

		2	3	>= 4
10-shot	Temporal	31.25	25.24	24.55
	Causal	17.39	9.52	6.52
	Subevent	0	16.0	0

Table 5: GPT-3.5 performance (F-1 score) using the Iterative Prediction 10-shot prompt on data groups with different event densities (measured by # of events within the same sentence).

We investigate the impact of event density on model performance by examining the predictions of GPT-3.5 on the 10 validation documents. We only consider event pairs within the same sentence. Event density is measured as the number of event and TIMEX mentions appeared in one sentence. As shown in Table 5, the performance of GPT-3.5 on temporal and causal relations decreases quickly as the event density increases, indicating GPT-3.5 struggles to capture event relations when the complexity of the context increases.

6 Conclusion

In this study, we systematically evaluated the effectiveness of LLMs in performing discourse-level ERE tasks featuring lengthy documents and intricate relations. Our experiments using multiple prompt patterns uncover a noteworthy underperformance of LLMs when compared to the baseline established through supervised learning. Even with supervised fine-tuning, LLMs like LLaMA-2 still underperform the much smaller supervised baseline model when trained on the same amount of data. Furthermore, our quantitative and qualitative analyses revealed that LLMs face challenges in obeying transitivity rules, capturing inter-sentence

relations or relations with a long distance, as well as comprehending context with dense event mentions. For future work, we will further investigate these challenges and develop methods for enabling LLMs to better address some of these issues in event relation extraction.

Limitation

Although we tried several different prompt patterns, there is still a chance that there exists better prompt to be used to assist GPT to solve the ERE task better. Meanwhile, OpenAI constantly update the GPT models that can be accessed throught the API, making it hard to reproduce the results if older models are deprecated. While OpenAI has offered preliminary introductions to various versions of GPT models, the specific implementation details remain obscure. This opacity hampers thorough analysis of why distinct versions of GPT models exhibit varying performance levels and how each data set and training technique influences models' performance. Finally, OpenAI API is a paid service, conducting experiments can get expensive depending on the task and design of the evaluation, making it not accessible to larger community. We are also limited by the cost and response time of OpenAI API.

For LLaMA-2 models, larger models (13B and 70B) may have better performance, but we leave the thorough study of LLaMA-2 models for potential future works.

Ethics and Broader Impact

We are aware that such study is very expensive and not very accessible to some researchers in the NLP community as OpenAI API is a paid service and is restricted in many countries. Not all researchers in our community can afford thousands of dollars or even more to run such experiments. Experiments on exclusive models or API may further detriment the inclusiveness of NLP community. Therefore, we hope our work can provide insights to readers with limited resources and inspire them in other works. However, by no means that we are advocating the NLP community to include closed-source LLMs as the baseline for any of the future work as studying the performance and behavior of closed-source models can be extremely difficult. We aim for our work to serve as a valuable resource for readers, helping them make decisions as they leverage LLMs for complex ERE tasks at discourse-level.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation via the award IIS-1942918. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4780–4790, Florence, Italy. Association for Computational Linguistics.

Robert L. Logan IV au2, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Long Bai, Saiping Guan, Jiafeng Guo, Zixuan Li, Xiaolong Jin, and Xueqi Cheng. 2021. Integrating deep event-level and script-level information for script event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9869–9878, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Ali Borji. 2023. A categorical archive of chatgpt failures.

David Van Buren. 2023. Guided scenarios with simulated expert personae: a remarkable strategy to perform cognitive work.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2018. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 485–495, Melbourne, Australia. Association for Computational Linguistics.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain. Association for Computational Linguistics.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023a. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023b. Exploring the feasibility of chatgpt for event extraction.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. 2023. Yes but.. can chatgpt identify entities in historical documents?

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Zhilei Hu, Zixuan Li, Daozhu Xu, Long Bai, Cheng Jin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2023. Protoem: A prototype-enhanced matching framework for event relation extraction.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2019. Question answering as global reasoning over semantic abstractions.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering.

William J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zhengzhong Liu, Jun Araki, Eduard Hovy, and Teruko Mitamura. 2014. Supervised within-document event coreference using information propagation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4539–4544, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jing Lu and Vincent Ng. 2021. Constrained multi-task learning for event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4504–4514, Online. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and Perception*. Harvard University Press, Cambridge, MA and London, England.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 415–424, New York, NY, USA. Association for Computing Machinery.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Marta Recasens and Edward Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Cathy Shyr, Yan Hu, Paul A. Harris, and Hua Xu. 2023. Identifying and extracting rare disease phenotypes with large language models.

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

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Kangda Wei, Sayan Ghosh, Rakesh Menon, and Shashank Srivastava. 2023b. Leveraging multiple teachers for test-time adaptation of language-guided classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7068–7088, Singapore. Association for Computational Linguistics.

Kangda Wei, Dawn Lawrie, Benjamin Van Durme, Yunmo Chen, and Orion Weller. 2023c. When do decompositions help for machine reading? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3599–3606, Singapore. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023d. Zero-shot information extraction via chatting with chatgpt.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4004–4010. International Joint Conferences on Artificial Intelligence Organization. Main track.

Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. ACM.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.

Appendix

A Hyperparameter and Compute Detail

We use the default hyperparameters for *gpt-3.5-turbo-16k* and *gpt-4-1106-preview* with the temperature set to 1, top_p set to 1, frequency penalty set to 0, and presence penalty set to 0. It takes around 48 - 60 hours to run though the test set of MAVEN-ERE for inference with *gpt-3.5-turbo-16k*. For LLaMA-2, we set temperature to 0.6, and top_p to 0.9, following the llama-recipes[8] GitHub repository maintained by Meta. It takes around 36 - 48 hours to run though the test set of MAVEN-ERE for inference with *Llama-2-7b-chat-hf*.

For the baseline model, we train the model following the hyperparameters in Wang et al. (2022). We train the model with the learning rate sets to $1e-5$ for the RoBERTa model, the learning rate sets to $1e-5$ for the classification head, and the batch size sets to 4. For coreference, temporal, and causal resolutions, we train the model for 50 epochs. For subevent resolution, we train the model for 20 epochs. The training and inference time in total varies from 30 minutes to 2.5 hours depending on the relation type.

For SFT with *Llama-2-7b-chat-hf*, we fine-tune the model for 3 epochs, with a learning rate of $2e-4$, weght decay of 0.001 for AdamW optimizer. 4-bit quantization is used to save memory space, and LoRA with 64 attention dimension, and 0.1 dropout rate is used for speeding up the training process.

B Models' Versions

The *gpt-3.5-turbo-16k* currently points to *gpt-3.5-turbo-0613*, which is a snapshot of *gpt-3.5-turbo* from June 13th 2023 and will be deprecated on June 13, 2024, according to the OpenAI website https://platform.openai.com/docs/models. All our final experiment runs were conducted in a relatively focused period of time, in November 2023, so the model we used is the *gpt-3.5-turbo-0613*.

The *gpt-4-1106-preview* was released on November 6th, 2023, and has knowledge of world events up to April 2023, according to the OpenAI website https://platform.openai.com/docs/models. All our final experiment runs were conducted in a relatively focused period of time, in January 2024.

[8] https://github.com/meta-llama/llama-recipes/tree/main

The LLaMA-2 model used in this paer is *Llama-2-7b-chat-hf*, which can be accessed through Huggingface https://huggingface.co/meta-llama/Llama-2-7b-chat-hf.

C GPT-3.5 Experiments Cost

We use the *gpt-3.5-turbo-16k* model from OpenAI API. The cost for *gpt-3.5-turbo-16k* is $0.001 per 1k tokens for input and $0.002 per 1k tokens for output. We run through the MAVEN-ERE test set (857 documents) for 5 times since we run through the whole test set once for each of the whole doc, 1-shot, 2-shot, 5-shot, and 10-shot prompt. The total cost is around $1650, resulting $330 on average for each different prompt pattern used, $0.385 on average for annotating a document.

D Prompt Patterns

Here, we provide some examples for the whole doc and n-shot prompt described in Sec 3.2. The system prompts for coreference, temporal, causal and subevent relations are shown in Table 6. We show an example of whole doc prompt for coreference and temporal relations in Table 7 , and an example of 2-shot prompt for coreference and temporal relations in Table 8. Causal and subevent relations follow the same pattern as temporal relation.

We also show some examples for the two other prompt patterns: Bulk Prediction and Event Ranking. Examples for these two prompt patterns are shown in Table 9. We don't choose to use the Bulk Prediction prompt is because the performance is overall the worst comparing to other prompts when testing on the first 10 documents of the validation set. Event Ranking has relatively good performance on the first 10 validation document but require much more number of queries. We estimate a total cost of $3,200 if use the event ranking prompt to run through the whole test set once. Event Ranking prompt is also very time-consuming with an estimated 400 hours to run through the test set. Since using event ranking prompt is both financially and time-wise impossible, we don't choose to use it.

E Detailed report of GPT-3.5, GPT-4 and LLaMA-2 Performance Over Validation Set

Here, we report the performance (precision, recall, and F-1 scores) of GPT-3.5 on the first 10 validation documents using the Iterative Prediction prompt patterns in Table 10. We report the performance (precision, recall, and F-1 scores) of GPT-4 on the

Relation Type	System
Coreference	You are an annotator for the MAVEN-ERE dataset. Your task is to extract event coreference relations between event mentions from given documents, where all event and TIMEX mentions are given in [ ]. Imitate the given example to find coreference relations between event mentions. The last sentence of the context is not annotated. You should find all the relations among the new mentions in the last sentence with mentions in all previous sentences. Predict the relations in this format: Event_1 COREFERENCE Event_0; SHIFT; means moving on to the next sentence. Always add SHIFT; at the end of prediction.
Temporal	You are an annotator for the MAVEN-ERE dataset. Your task is to extract temporal relations between event mentions from given documents, where all event and TIMEX mentions are given in [ ]. There are 6 types of temporal relations: BEFORE, CONTAINS, OVERLAP, BEGINS-ON, ENDS-ON, and SIMULTANEOUS. Imitate the given example to find temporal relations between event and TIMEX mentions. The last sentence of the context is not annotated. You should find all the relations among the new mentions in the last sentence with mentions in all previous sentences. Predict the relations in this format: Event_1 BEFORE Event_0; SHIFT; means moving on to the next sentence. Always add SHIFT; at the end of prediction.
Causal	You are an annotator for the MAVEN-ERE dataset. Your task is to extract causal relations between event mentions from given documents, where all event and TIMEX mentions are given in [ ]. There are 2 types of causal relations: CAUSE, and PRECONDITION. Imitate the given example to find causal relations between event and TIMEX mentions. The last sentence of the context is not annotated. You should find all the relations among the new mentions in the last sentence with mentions in all previous sentences. Predict the relations in this format: Event_1 CAUSE Event_0; SHIFT; means moving on to the next sentence. Always add SHIFT; at the end of prediction.
Subevent	You are an annotator for the MAVEN-ERE dataset. Your task is to extract subevent relations between event mentions from given documents, where all event and TIMEX mentions are given in [ ]. Imitate the given example to find subevent relations between event and TIMEX mentions. The last sentence of the context is not annotated. You should find all the relations among the new mentions in the last sentence with mentions in all previous sentences. Predict the relations in this format: Event_1 SUBEVENT Event_0; SHIFT; means moving on to the next sentence. Always add SHIFT; at the end of prediction.

Table 6: System prompts for causal and subevent relations.

first 10 validation documents using the Iterative Prediction prompt patterns in Table 11. we report the performance (precision, recall, and F-1 scores) of LLaMA-2 on the first 10 validation documents using the Iterative Prediction prompt patterns in Table 12.

F Precision and Recall Results for LLaMA-2 SFT Experiment

We show the Precision scores of coreference, temporal, causal, and subevent relations for the size experiment discussed in Section 4 in Figure 4, and we also show the Recall scores of coreference, temporal, causal, and subevent relations for the size experiment discussed in Section 4 in Figure 5

Relation type	**Coreference**	**Temporal**
Prompt	The [0 Expedition Event_0] of the Thousand (Italian " Spedizione dei Mille ") was an event of the Italian Risorgimento that [took place Event_1] in [1860 TIMEX_0]. a corps of volunteers led by giuseppe garibaldi [sailed Event_2] from quarto, near genoa (now quarto dei mille) and [landed Event_3] in marsala, sicily, in order to [1 conquer Event_4] the kingdom of the two sicilies, [ruled Event_5] by the house of bourbon-two sicilies. The project was an ambitious and risky [0 venture Event_6] [aiming Event_7] to [1 conquer Event_8], with a thousand men, a kingdom with a larger regular army and a more powerful navy. / Event_9 COREFERENCE 0; Event_13 COREFERENCE 1; SHIFT; The King David Hotel [0 bombing Event_0] was a terrorist [0 attack Event_1] [carried out Event_2] on [Monday, July 22, 1946 TIMEX_0], by the militant right-wing Zionist underground organization the Irgun on the British administrative headquarters for Palestine, which was housed in the southern wing of the King David Hotel in Jerusalem during the Jewish insurgency in Mandatory Palestine. 91 people of various nationalities were [killed Event_3], and 46 were [injured Event_4]. the hotel was the site of the central offices of the british mandatory authorities of palestine, principally the secretariat of the government of palestine and the headquarters of the british armed forces in palestine and transjordan. When [planned Event_5], the [attack Event_6] had the [approval Event_7] of the Haganah, the principal Jewish paramilitary group in Palestine, though, unbeknownst to the Irgun, this had been [cancelled Event_8] by the time the [operation Event_9] was [carried out Event_10].	The [Expedition Event_0 TIMEX_0 CONTAINS Event_0;Event_5 OVERLAP Event_0;] of the Thousand (Italian 'Spedizione dei Mille') was an event of the Italian Risorgimento that [took place Event_1 TIMEX_0 CONTAINS Event_1;Event_5 OVERLAP Event_1;Event_0 SIMULTANEOUS Event_1;] in [1860 TIMEX_0 Event_5 OVERLAP TIMEX_0;]. A corps of volunteers led by giuseppe garibaldi [sailed Event_2] from quarto, near genoa (now quarto dei mille) and [landed Event_3] in marsala, sicily, in order to [conquer Event_4] the kingdom of the two sicilies, [ruled Event_5] by the house of bourbon-two sicilies. / Event_0 CONTAINS Event_2; Event_1 CONTAINS Event_2; TIMEX_0 CONTAINS Event_2; Event_0 CONTAINS Event_3; Event_1 CONTAINS Event_3; ... Event_5 CONTAINS Event_3; SHIFT; The Cherry Valley massacre was an attack by British and Iroquois forces on a fort and the village of Cherry Valley in eastern New York on [November 11, 1778 TIMEX_0 TIMEX_1 CONTAINS TIMEX_0;], during [the American Revolutionary War TIMEX_1]. It has been [described Event_0] as one of the most horrific frontier massacres of the war.

Table 7: An example of the whole doc prompt on coreference and temporal relations. Causal and subevent relations follow the same pattern as temporal relation. The sentence in red is the sentence being queried.

Relation type	Coreference	Temporal
Prompt	The Cherry Valley massacre was an attack by British and Iroquois forces on a fort and the village of Cherry Valley in eastern New York on [November 11, 1778 TIMEX_0], during [the American Revolutionary War TIMEX_1]. / SHIFT; The Cherry Valley massacre was an attack by British and Iroquois forces on a fort and the village of Cherry Valley in eastern New York on [November 11, 1778 TIMEX_0], during [the American Revolutionary War TIMEX_1]. It has been [described Event_0] as one of the most horrific frontier massacres of the war. / SHIFT; The King David Hotel [bombing Event_0] was a terrorist [attack Event_1] [carried out Event_2] on [Monday, July 22, 1946 TIMEX_0], by the militant right-wing Zionist underground organization the Irgun on the British administrative headquarters for Palestine, which was housed in the southern wing of the King David Hotel in Jerusalem during the Jewish insurgency in Mandatory Palestine. / Event_1 COREFERENCE Event_0; SHIFT; The King David Hotel [0 bombing Event_0] was a terrorist [0 attack Event_1] [carried out Event_2] on [Monday, July 22, 1946 TIMEX_0], by the militant right-wing Zionist underground organization the Irgun on the British administrative headquarters for Palestine, which was housed in the southern wing of the King David Hotel in Jerusalem during the Jewish insurgency in Mandatory Palestine. 91 people of various nationalities were [killed Event_3], and 46 were [injured Event_4]. / SHIFT; The [Battle Event_0] of Orthez ([27 February 1814 TIMEX_0]) saw the Anglo-Portuguese Army under Field Marshal Arthur Wellesley, Marquess of Wellington [attack Event_1] an Imperial French army [led Event_2] by Marshal Nicolas Soult in southern France. The outnumbered French [repelled Event_3] several Allied [assaults Event_4] on their right flank, but their center and left flank were [overcome Event_5] and Soult was compelled to [retreat Event_6].	The Cherry Valley massacre was an attack by British and Iroquois forces on a fort and the village of Cherry Valley in eastern New York on [November 11, 1778 TIMEX_0], during [the American Revolutionary War TIMEX_1]. / TIMEX_1 CONTAINS TIMEX_0; SHIFT; The Cherry Valley massacre was an attack by British and Iroquois forces on a fort and the village of Cherry Valley in eastern New York on [November 11, 1778 TIMEX_0 TIMEX_1 CONTAINS TIMEX_0;], during [the American Revolutionary War TIMEX_1]. It has been [described Event_0] as one of the most horrific frontier massacres of the war. / SHIFT; The United States occupation of Nicaragua from [1912 TIMEX_0] to [1933 TIMEX_1] was part of the Banana Wars, when the US military intervened in various Latin American countries from [1898 TIMEX_2] to [1934 TIMEX_3]. / TIMEX_0 BEFORE TIMEX_1; TIMEX_2 BEFORE TIMEX_0; TIMEX_2 BEFORE TIMEX_1; TIMEX_0 BEFORE TIMEX_3; TIMEX_1 BEFORE TIMEX_3; TIMEX_2 BEFORE TIMEX_3; SHIFT; The United States occupation of Nicaragua from [1912 TIMEX_0 TIMEX_2 BEFORE TIMEX_0;] to [1933 TIMEX_1 TIMEX_0 BEFORE TIMEX_1;TIMEX_2 BEFORE TIMEX_1;Event_0 BEFORE TIMEX_1;TIMEX_4 BEFORE TIMEX_1;] was part of the Banana Wars, when the US military intervened in various Latin American countries from [1898 TIMEX_2] to [1934 TIMEX_3 TIMEX_0 BEFORE TIMEX_3;TIMEX_1 BEFORE TIMEX_3;TIMEX_2 BEFORE TIMEX_3;Event_0 BEFORE TIMEX_3;TIMEX_4 BEFORE TIMEX_3;]. The formal occupation [began Event_0] in [1912 TIMEX_4], even though there were various other assaults by the U.S. in Nicaragua throughout this period. / TIMEX_0 BEFORE Event_0; Event_0 BEFORE TIMEX_1; TIMEX_2 BEFORE Event_0; Event_0 BEFORE TIMEX_3; TIMEX_0 BEFORE TIMEX_4; TIMEX_4 BEFORE TIMEX_1; TIMEX_2 BEFORE TIMEX_4; TIMEX_4 BEFORE TIMEX_3; TIMEX_4 CONTAINS Event_0; SHIFT; The [Battle Event_0 TIMEX_0 SIMULTANEOUS Event_0;] of Malacca ([2 August 1640 – 14 January 1641 TIMEX_0]) was a successful attempt by the Dutch to [capture Event_1 Event_0 BEFORE Event_1;TIMEX_0 BEFORE Event_1;] Malacca from the Portuguese. In [the early 17th century TIMEX_1], the Dutch East India Company (Verenigde Oostindische Compagnie) [began Event_2] the campaign to [destroy Event_3] Portuguese power in the East.

Table 8: An example of the 2-shot prompt on coreference and temporal relations. Causal and subevent relations follow the same pattern as temporal relation. The sentence in red is the sentence being queried.

Prompt Method	Bulk Prediction	Event Ranking
System	You are an annotator for the MAVEN-ERE dataset. Your task is to extract event coreference, temporal, causal, and subevent relations between event and TIMEX mentions from given documents, where all event and TIMEX mentions are given. Coreference and subevent relations are binary. For temporal relations, there are 6 types: ... For causal relations, there are 2 types: ... Note that the order of the events matter. SIMULTANEOUS and BEGINS-ON are bidirectional relations. If there is no relations, return an empty array. You should always finish the answer instead of using '...'.	You are an annotator for the MAVEN-ERE dataset. Your task is to extract event coreference, temporal, causal, and subevent relations between event and TIMEX mentions from given documents, where all event and TIMEX mentions are given in [] after triggering words. All predictions should be an array with elements being EVENT and TIMEX mentions given in [] from document.
Prompt	This is the document: The Expedition [Event_0] of the Thousand (Italian 'Spedizione dei Mille') was ... in 1860 [TIMEX_0]... distribution [Event_26] and the [Event_27] end of oppression. What are the temporal relations? / {BEFORE: [[Event_17, Event_0], [Event_19, Event_0]...], CONTAINS: [[Event_7, Event_0], [TIMEX_0, Event_0]...], OVERLAP: [[Event_5, Event_0], [Event_5, Event_1]...], ... SIMULTANEOUS: [[Event_0, Event_1], [Event_15, Event_1]...]} This is the document: The Cherry Valley massacre was an attack ... on November 11, 1778 [TIMEX_0], during ... leading to the 1779 [TIMEX_3] Sullivan Expedition which drove [Event_14] the Iroquois out of western New York. What are the temporal relations?	This is the document: The Expedition [Event_0] of the Thousand (Italian 'Spedizione dei Mille') was ... in 1860 [TIMEX_0]... distribution [Event_26] and the [Event_27] end of oppression. List all mentions happened BEFORE [Event_0]. If there is no relations, return an empty array. / [Event_17, Event_19, Event_20, Event_21] This is the document: The Cherry Valley massacre was an attack ... on November 11, 1778 [TIMEX_0], during ... leading to the 1779 [TIMEX_3] Sullivan Expedition which drove [Event_14] the Iroquois out of western New York. List all mentions happened BEFORE [Event_0]. If there is no relations, return an empty array.

Table 9: Other prompt patterns that are tried.

	MUC			B^3			$CEAF_e$			BLANC		
Prompt	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
Bulk Prediction whole doc (1-shot)	3.3	14.3	5.3	77.7	95.1	85.5	87.4	71.0	78.3	51.4	53.6	51.9
Iterative Prediction whole doc	9.5	28.6	14.3	85.3	95.8	90.2	92.8	82.4	87.3	52.6	59.8	53.8
1-shot	2.4	7.1	3.6	84.6	94.9	89.4	91.9	82.0	86.6	50.2	50.9	50.0
2-shot	**10.6**	**35.7**	**16.4**	83.2	**96.0**	89.2	93.0	80.8	86.5	52.3	61.6	53.4
5-shot	7.7	21.4	11.3	**86.0**	95.6	**90.5**	92.5	83.2	87.6	**52.9**	**61.9**	**54.3**
10-shot	5.1	14.3	7.6	85.8	95.3	90.3	**93.2**	**83.9**	**88.3**	50.9	53.3	51.2
Event Ranking whole doc (1-shot)	4.4	14.3	6.7	83.3	95.1	88.8	88.9	77.5	82.8	51.9	53.8	52.5

	Temporal			Causal			Subevent			Overall		
Prompt	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
Bulk Prediction whole doc (1-shot)	7.2	2.3	3.4	2.8	2.4	2.6	3.0	**9.8**	4.6	17.0	18.3	16.5
Iterative Prediction whole doc	19.9	8.6	12.0	2.2	4.2	2.9	2.1	7.3	3.3	21.1	21.7	19.9
1-shot	18.0	10.4	13.2	5.0	7.7	6.1	2.8	**9.8**	4.3	20.8	21.7	20.3
2-shot	18.5	7.1	10.2	4.6	6.0	5.2	3.2	7.3	4.4	21.5	22.2	20.3
5-shot	16.3	8.0	10.8	3.9	4.8	4.3	0.0	0.0	0.0	20.0	19.6	19.0
10-shot	**22.1**	10.6	14.3	**6.2**	**10.1**	**7.7**	2.0	**9.8**	3.3	**22.3**	23.1	21.2
Event Ranking whole doc (1-shot)	15.8	**34.1**	**21.6**	4.8	7.6	5.9	**4.4**	7.9	**5.7**	20.5	**27.4**	**22.7**

Table 10: Event coreference, temproal, causal, and subevent resolution performances of *gpt-3.5-turbo-16k* on `MAVEN-ERE` using different prompt patterns on the first 10 documents of the validation set. The numbers in **bold** indicate the best performance across different prompt patterns for *gpt-3.5-turbo-16k*. Notice that the Event Ranking 1-shot prompt has the best overall performance and the Bulk Prediction 1-shot prompt has the worst overall performance.

Prompt	MUC Precision	MUC Recall	MUC F-1	B^3 Precision	B^3 Recall	B^3 F-1	$CEAF_e$ Precision	$CEAF_e$ Recall	$CEAF_e$ F-1	BLANC Precision	BLANC Recall	BLANC F-1
Iterative Prediction												
whole doc	11.5	21.4	15.0	91.3	95.5	93.3	93.0	88.6	90.7	52.5	55.9	53.4
1-shot	16.7	21.4	18.8	94.3	95.5	94.9	94.2	92.7	93.5	58.9	58.5	58.7
2-shot	27.3	21.4	24.0	97.0	95.5	96.2	93.5	94.6	94.1	66.4	58.6	61.3
5-shot	25.0	21.4	23.1	96.6	95.3	96.1	94.0	94.7	94.4	61.5	58.5	59.8
10-shot	14.3	16.7	15.4	94.6	95.5	95.2	93.7	92.9	93.3	55.3	54.8	55.0

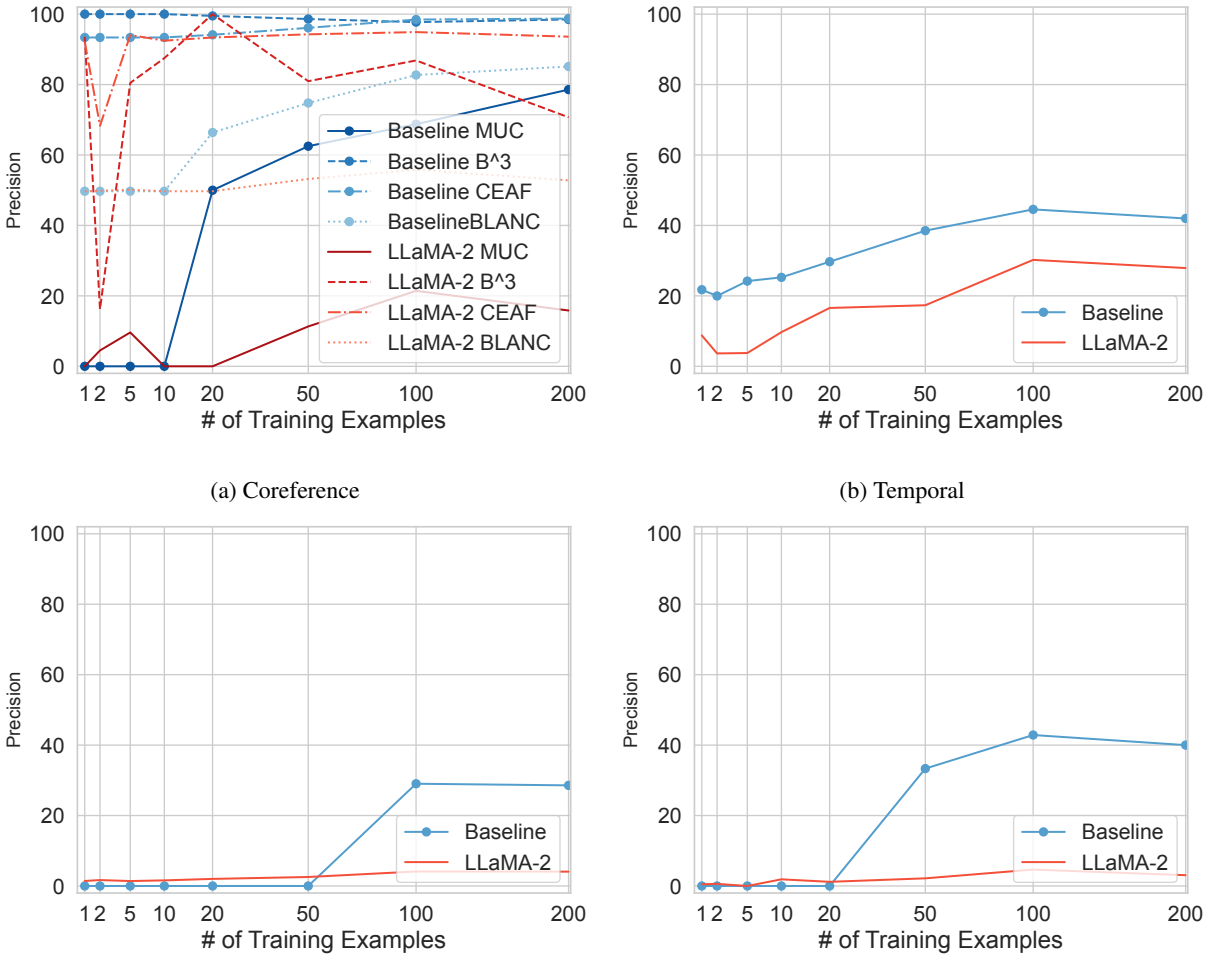
Prompt	Temporal Precision	Temporal Recall	Temporal F-1	Causal Precision	Causal Recall	Causal F-1	Subevent Precision	Subevent Recall	Subevent F-1	Overall Precision	Overall Recall	Overall F-1
Iterative Prediction												
whole doc	25.8	19.0	21.9	3.2	4.8	3.8	0	0	0	22.8	20.9	22.2
1-shot	19.0	10.5	13.6	6.0	6.0	6.0	0	0	0	22.8.	20.9	21.5
2-shot	21.1	8.4	12.1	5.4	4.2	4.7	4.3	4.9	4.6	25.5	21.3	22.6
5-shot	21.6	11.6	15.1	10.4	6.0	7.6	1.9	2.4	2.1	25.8	21.9	23.3
10-shot	25.2	12.7	16.9	12.2	9.4	10.6	0	0	0	25.5	21.8	23.1

Table 11: Event coreference, temporal, causal, and subevent resolution performances of *gpt-4-1106-preview* on `MAVEN-ERE` using Iterative Prediction prompt patterns on the first 10 documents of the validation set.

Prompt	MUC Precision	MUC Recall	MUC F-1	B^3 Precision	B^3 Recall	B^3 F-1	$CEAF_e$ Precision	$CEAF_e$ Recall	$CEAF_e$ F-1	BLANC Precision	BLANC Recall	BLANC F-1
Iterative Prediction												
whole doc	4.6	7.1	5.6	91.9	94.9	93.3	92.5	90.0	91.0	50.9	51.6	51.1
1-shot	0	0	0	100.0	94.7	97.3	93.4	98.6	95.9	49.7	50.0	49.9
2-shot	0	0	0	100.0	94.7	97.3	93.4	98.6	95.9	49.7	50.0	49.9
5-shot	0	0	0	100.0	94.7	97.3	93.4	98.6	95.9	49.7	50.0	49.9

Prompt	Temporal Precision	Temporal Recall	Temporal F-1	Causal Precision	Causal Recall	Causal F-1	Subevent Precision	Subevent Recall	Subevent F-1	Overall Precision	Overall Recall	Overall F-1
Iterative Prediction												
whole doc	16.7	3.5	5.7	5.8	12.5	7.9	3.4	26.8	6.1	22.8	20.9	20.0
1-shot	8.4	1.9	3.0	8.7	1.2	2.1	0	0	0	22.8	20.9	16.5
2-shot	16.4	3.1	5.2	9.1	0.6	1.1	0	0	0	25.5	21.3	16.8
5-shot	13.5	2.7	4.5	0	0	0	0	0	0	25.8	21.9	16.3

Table 12: Event coreference, temporal, causal, and subevent resolution performances of *LLaMA-2* on `mavenere` using Iterative Prediction prompt patterns on the first 10 documents of the validation set.



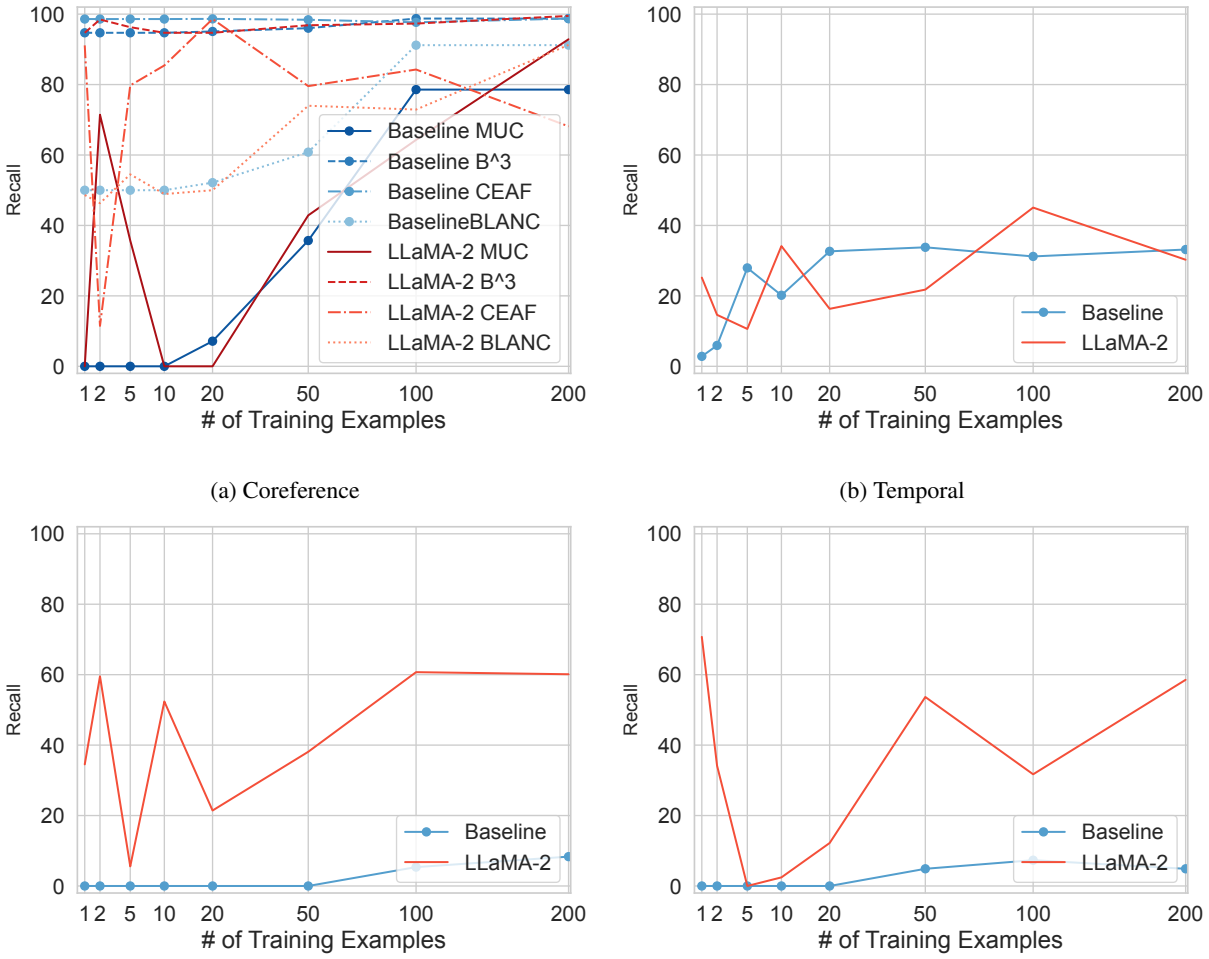
(a) Coreference

(b) Temporal

(c) Causal

(d) Subevent

Figure 4: Precision scores for coreference, temporal, causal, and subevent relations. The blue lines represent the baseline's performance trained with different numbers of documents. The red lines represent the performance of LLaMA-2 fine-tuned with different numbers of documents.



(a) Coreference

(b) Temporal

(c) Causal

(d) Subevent

Figure 5: Recall scores for coreference, temporal, causal, and subevent relations. The blue lines represent the baseline's performance trained with different numbers of documents. The red lines represent the performance of LLaMA-2 fine-tuned with different numbers of documents.