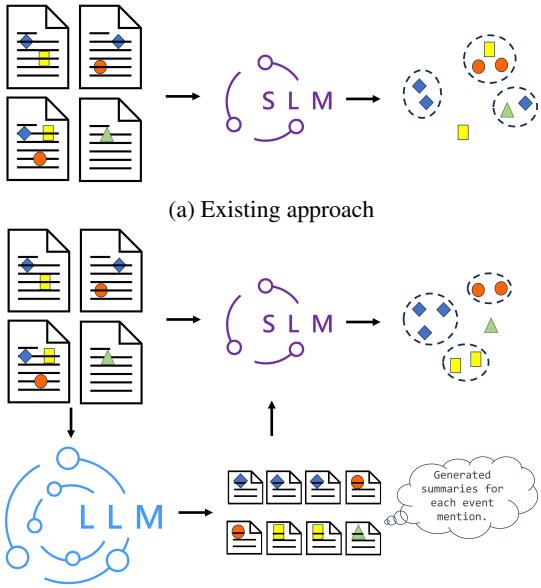
Synergetic Event Understanding: A Collaborative Approach to Cross-Document Event Coreference Resolution with Large Language Models

Anonymous ACL submission

Abstract

Cross-document event coreference resolution (CDECR) involves clustering event mentions across multiple documents that refer to the same real-world events. Existing approaches utilize fine-tuning of small language models (SLMs) like BERT to address the compatibility among the contexts of event mentions. However, due to the complexity and diversity of contexts, these models are prone to learning simple co-occurrences. Recently, large language models (LLMs) like ChatGPT have demonstrated impressive contextual understanding, yet they encounter challenges in adapting to specific information extraction (IE) tasks. In this paper, we propose a collaborative approach for CDECR, leveraging the capabilities of both a universally capable LLM and a task-specific SLM. The collaborative strategy begins with the LLM accurately and comprehensively summarizing events through prompting. Then, the SLM refines its learning of event representations based on these insights during fine-tuning. Experimental results demonstrate that our approach surpasses the performance of both the large and small language models individually, forming a complementary advantage. Across various datasets, our approach achieves state-of-the-art performance, underscoring its effectiveness in diverse scenarios.



(a) Existing approach

(b) Our approach

Figure 1: Models for cross-document event coreference resolution, where the input comprises event mentions from different documents, and the output consists of event clusters formed by coreferential mentions, which are visually represented by icons sharing the same color and shape.

1 Introduction

Event coreference resolution is a useful task in information extraction (Lu and Ng, 2018). This is crucial for achieving a more comprehensive understanding of intricate narratives and facilitating knowledge extraction from diverse textual sources. The coreference of events typically relies on a thorough understanding of document-level context (Minh Tran et al., 2021; Kriman and Ji, 2021; Xu et al., 2022). Cross-document event coreference (Lee et al., 2012), involving the comparison of event mentions from different documents, presents additional challenges. On one hand, distinct events in different documents may be portrayed in a very similar manner, especially for events of the same type (challenge 1). On the other hand, the portrayal of the identical event may vary significantly across different documents (challenge 2). The model is required to grasp comparable coreference evidence from varied contexts and make judgments based on it (refer to the examples in Table 4 and 10 for better illustration).

Existing work (Held et al., 2021; Yu et al., 2022) attempts to address CDECR based on fine-tuning small language models (SLMs) [1], as shown in Fig-

[1] In this work, SLM refers to pre-trained language models with relatively fewer parameters, which are more cost-effective for fine-tuning on specific tasks, such as BERT and RoBERTa.

ure 1a. However, the complexity and diversity of the context make it prone to learning pseudo-features by capturing simple co-occurrences rather than genuinely coreference-related terms, including contextual words, entity mentions and other event mentions associated with the given event mention. Supporting this observation, CDECR remains a significant challenge for SLMs, as evidenced by achieving only around 70% CoNLL F1 score on the FCC dataset (Bugert et al., 2021).

Recent advancements in LLMs have significantly advanced the field of NLP, enabling the effective resolution for tasks like machine translation (Jiao et al., 2023) and text summarization (Bang et al., 2023), with just a few demonstrations. However, when it comes to information extraction (IE) tasks, LLMs encounter challenges in task-specific adaptation. Specifically, LLMs struggle to achieve the same level of accuracy as supervised SLMs because a small number of demonstrations cannot comprehensively cover the complex annotation guidelines of these tasks (Han et al., 2023; Li et al., 2023a). Moreover, the inherent nature of the CDECR task, which involves processing multiple documents, imposes enhanced demands on understanding the lengthy context in the demonstrations. [2] Instead of directly predicting CDECR structures, the relative strength of LLMs can enhance the generic understanding of individual documents, particularly the inherent meaning of diverse event mentions, which is complementary to the advantage of SLMs in understanding structures with thorough fine-tuning.

To leverage the relative strengths of LLMs and SLMs, we propose a collaborative approach, as shown in Figure 1b. First, we use the LLM to summarize event mentions from different documents. Then we feed these insights to the SLM to enhance its understanding of event mentions, enabling it to make coreference judgments based on more focused contexts. For the LLM summarization, we design a two-step workflow with separate generic prompts to guide its comprehension of the context of each mention, instead of task-specific in-context learning or fine-tuning. For the SLM, we employ joint representation learning to integrate the original document and the generated summary.

We conduct experiments on three datasets of CDECR, and the results demonstrate that our collaborative approach, as compared to methods solely relying on the LLM or SLM, exhibits significant improvements. Across all three datasets (ECB+, GVC, and FCC), our approach achieves state-of-the-art results, with increases of 1%, 2.7%, and 7% in CoNLL F1, respectively (averaged over three independent experiments for each dataset). Through analysis, it is demonstrated that our approach can more thoroughly address the aforementioned challenge 1 of similarly portrayed contexts, making a substantial contribution to performance improvement. Our code and data are released on XXX.

To the best of our knowledge, we are the first to propose a collaborative approach that leverages the universal capabilities of LLMs to address CDECR, achieving superior performance compared to the state-of-the-art baseline.

2 Related Work

CDECR Early work addresses CDECR by employing machine learning methods with manually designed features (Bejan and Harabagiu, 2010; Yang et al., 2015; Vossen and Cybulska, 2018; Bugert et al., 2021). Recent neural approaches have utilized SLMs to encode event mentions, obtaining their embeddings for supervised coreference resolution. Initial efforts involve encoding at sentence level and fusing the embeddings of mentions and the incomplete arguments extracted by SRL as the representation of mentions (Barhom et al., 2019; Zeng et al., 2020; Allaway et al., 2021; Yu et al., 2022). Subsequent work incorporates extensive context directly into encoding, leading to noticeable improvements (Caciularu et al., 2021; Cattan et al., 2021; Held et al., 2021; Hsu and Horwood, 2022; Ahmed et al., 2023). More recently, Chen et al. (2023) and Ravi et al. (2023) further establish connections between event mentions using a discourse rhetorical structure constructor and a GPT-3 model fine-tuned with additional manually annotated data specialized in temporal reasoning, respectively. In comparison to existing work, we are the first to establish comprehensive connections between event mentions and their corresponding contextual elements, including contextual words, entity mentions, and other event mentions, by leveraging the intrinsic knowledge and out-of-the-box context comprehension ability of LLMs.

LLM for IE Several recent studiess (Ma et al., 2023; Li et al., 2023a; Han et al., 2023; Yuan et al., 2023; Gao et al., 2023; Wei et al., 2023; Xie et al.,

[2] On average, each instance of demonstration in the ECB+ dataset contains close to 15k tokens.

2023; Li and Zhang, 2023; Xu et al., 2023; Wadhwa et al., 2023; Qi et al., 2023; Ling et al., 2023) have evaluated the performance of LLMs, predominantly ChatGPT, using in-context learning methods on various IE tasks. These investigations universally demonstrate that LLMs exhibit commendable performance in zero-shot and few-shot settings, yet there remains a substantial gap when compared to state-of-the-art supervised SLMs, with the performance gap widening for more complex tasks.

In addition, there are also methods directly using labeled data from IE tasks to fine-tune LLMs(Lu et al., 2022; Zhou et al., 2023; Wang et al., 2023; Sainz et al., 2023). In general, training on these larger-scale models, such as Code-LLaMA and Flan-T5, has yielded results comparable to supervised baselines and demonstrated improvements in zero-shot settings. However, when the training of LLMs does not result in significant performance gains, the training cost, compared to training SLMs, becomes less cost-effective.

Integration of LLM and SLM The integration of LLM and SLM is an emerging approach, with only a few explorations in complex IE tasks. Ma et al. (2023) prompts the LLM to rerank a small portion of difficult samples filtered by the supervised SLM and achieves promising improvements on various few-shot IE tasks. Their method is based on the observation that LLMs excel only at a small number of hard samples compared with SLMs. Wan et al. (2023) utilizes a fine-tuned SLM to retrieve demonstrations and further inject the reasoning logic generated by the LLM into the demonstration for prompting. Their method surpasses the supervised baseline on certain relation extraction datasets. However, it faces a drawback as LLM performance heavily depends on the quality of SLM-retrieved demonstrations, especially for NULL-type triples, where finding suitable demonstrations is inherently challenging. This hinders performance on complex datasets, such as ACE05. Additionally, inducing complex reasoning logic for each of the k-demonstrations incurs high costs for LLM, leading them to sample only a subset from the original test set of ACE05 and TACRED. Xu et al. (2023) and Li et al. (2023b) leverage LLMs for data enhancement in sentence and document-level relation extraction tasks, respectively. The gap between triples recognized by LLMs and those annotated under manually crafted rules introduces potential shifts in the data distribution, making the effectiveness in practical applications unclear.

Overall, the aforementioned integration methods have exhaustively attempted to adapt LLMs to specific tasks by prompting them to establish accurate connections with artificially defined labels. In contrast, our approach only requires LLMs to perform generic tasks, leveraging their inherent capabilities to assist specific tasks.

3 Method

We adopt the state-of-the-art method proposed by Held et al. (2021) as our baseline (Section 3.1), then summarize events using generic prompts for LLM (Section 3.2), and finally integrate the representations of events from both the summary and the baseline system (Section 3.3).

3.1 Task and Baseline

The goal of the CDECR task is to group coreferential event mentions across multiple documents into clusters. We formalize the task as follows:

Input: A corpus comprising multiple documents denoted by D, where $D = \{D_1, D_2, ..., D_{|D|}\}$, with $|D|$ representing the number of documents in the dataset. Let M represent all event mentions in the corpus, such that $M = \{m_{11}, m_{12}, \ldots, m_{ij}, \ldots, m_{|D|,k}\}$, where k denotes the number of event mentions in each document, and m_{ij} signifies the j-th event mention in document D_i.

Output: A set of clusters, denoted as C, where $C = \{C_1, C_2, ..., C_n\}$. For each cluster C_k, E_k represents all the event mentions contained in the cluster C_k, such that $E_k = \{e_{k1}, e_{k2}, \ldots, e_{kj}, \ldots, e_{kM}\}$, where M is the total number of event mentions in cluster C_k, and e_{kj} is the jth event mention in the cluster C_k.

Our baseline consists of two key modules for coreference clustering: candidate retrieval and pairwise classification. Both of these modules primarily involve using a RoBERTa (Liu et al., 2019) encoder to encode the context and obtain vector representations of event mentions, which can be seamlessly replaced by our collaborative approach. We formalize the encoding process as follows:

For each event mention m_{ij}, its vector representation can be obtained as:

$$h_{ij} = f_{\text{enc}}(m_{ij}, D_i) \quad (1)$$

Here f_{enc} is an encoder network used for encoding D_i and concatenating the representations of the boundary tokens of m_{ij}. The resulting representation h_{ij} is fed into the subsequent neural network.

Step 1	**News**: `[input document]` **Question 1:** In this news, given "`[mention 1]`" mentioned in the sentence "`[the sentence]`". Please elaborate `[mention 1]` in the context of the news article. Present the information in the following format: 'Elaboration: `[mention 1]` refers to <placeholder>'. **Question 2:** ...
Step 2	**News:** `[input document]` **Question 1:** In this news, given "`[mention 1]`" mentioned in the sentence "`[the sentence]`". Elaboration: `[output from step 1]`. Please further elaborate "`[mention 1]`" by providing details for entities in the elaboration utilizing coreference resolution. Provide any available or approximate dates in the news for reference, which can be inferred from the publication date of the news if available. Present the information in the following format: 'Elaboration: `[mention 1]` refers to <placeholder>'. **Question 2:** ...

Table 1: The two-step workflow for LLM summarization. Each prompt includes a document along with multiple event mentions. Step 2 takes the output from Step 1 as its input. The content to be filled is represented as `[content]`.

3.2 LLM Summarization

Summarizing events for CDECR poses a non-trivial challenge. Existing summarization methods are typically designed to provide a general overview of documents, making it difficult to extract information specific to certain types of events. This not only provides limited assistance for coreference but may also lead to the omission of crucial details. Furthermore, designing a summary template for each type of event is not only impractical in real-world applications [3] but also introduces bias, potentially causing LLMs to misinterpret or hallucinate information due to the inherent incompleteness of event information in documents.

To address various types of events and gather specific details from complex contexts, we design a two-step workflow to prompt the LLM, as shown in Table 1. The first step is responsible for extracting tailored information for different types of events in the context of the document. The second step aims to expand the details of the entities mentioned in the output of the first step, as entity details are often scattered throughout the document. In each step, we employ a straightforward prompt to accomplish a primary task. Our prompts adhere to the basic principle of faithfulness, avoiding additional interpretations to prevent semantic shifts. Compared to a synthesized single-step workflow, our two-step workflow guarantees that each step remains focused on its main objective, thereby preventing interference between the two steps, as illustrated by the analysis in Section 4.4.

In the first step, we instruct the LLM agent to "elaborate" an event mention, rather than the conventional instruction of "summarize". The term "elaborate" implies an explanatory behavior based on the concept of the mention words themselves, emphasizing the support of details from the document context. This suggests that LLMs can automatically select any relevant details from the context to support this explanation, including contextual words, entity mentions, and event mentions. This provides a standardized and feasible way to understand events, leveraging the LLM's intrinsic knowledge and contextual understanding capabilities without imposing complex rules for the LLM to adhere to.

In the second step, we prompt the LLM agent to use coreference resolution to aggregate detailed information about entities, as entity coreference is a more standardized task compared with event and performing it within a document reduces complexity. Additionally, we require the LLM to perform temporal reasoning based on the publication date of the document, further reducing ambiguity in coreference evidence comparison.

In both steps, we specify the generation format to ensure the consistency between the mention spans in summary and original document. This not only reduces the generation difficulty of LLM but also facilitates SLM in establishing the connection between the two during joint representation learning.

3.3 Integration into Final SLM

The SLM takes the original document and the generated summary as inputs. Through a direct joint representation learning technique, the new mention vector representation can be seamlessly integrated into the baseline.

Specifically, for the mention m_{ij}, let S_{ij} represents the generated summary, and $m_{ij}^{(s)}$ signifies

[3] Based on the rough statistics in our experiment, the ECB+ dataset contains over 400 event types.

the mention within it. By concatenating the original document D_i and the summary S_{ij}, a new document D'_i is formed. Let f'_{enc} denotes the new encoder network. It first encodes the new document D'_i, obtaining vector representations h_{ij} and $h^{(s)}_{ij}$ for m_{ij} and $m^{(s)}_{ij}$ respectively. These vectors are then concatenated to form the fused mention vector representation h'_{ij}, which can seamlessly substitute h_{ij} in the baseline for subsequent operations. The joint representation learning process can be represented as:

$$\begin{aligned} h'_{ij} &= f'_{\text{enc}}(\{e_{ij}, e^{(s)}_{ij}\}, D'_i) \\ &= \text{concat}\left(f_{\text{enc}}(\{e_{ij}, e^{(s)}_{ij}\}, D'_i)\right) \\ &= \text{concat}(h_{ij}, h^{(s)}_{ij}) \end{aligned}$$

Here $\{e_{ij}, e^{(s)}_{ij}\}$ denotes a set containing two elements, implying that vector representations for both e_{ij} and $e^{(s)}_{ij}$ can be derived using the same process as for a single element.

This integration method, which involves concatenating the original context and generated summary for joint representation learning, enables mutual learning of each other's context in the same attention space, thereby enhancing the understanding of genuinely coreference-related terms.

4 Experiments

4.1 Experimental Settings

Dataset We conduct experiments on three CDECR datasets: Event Coreference Bank Plus (ECB+) (Cybulska and Vossen, 2014), Gun Violence Corpus (GVC) (Vossen et al., 2018), and Football Coreference Corpus (FCC) (Bugert et al., 2021). The widely-used ECB+ dataset consists of news articles from various topics, including earthquakes, murders, acquisitions, etc. Each topic includes two similar subtopics, such as "6.1 earthquake Indonesia 2009" and "6.1 earthquake Indonesia 2013". This setup aligns with the challenge 1 mentioned in introduction, asking the model to distinguish similar events. Similarly, GVC and FCC, focusing on news incidents of gun violence and football tournaments, respectively, also have multiple subtopics under one overarching topic. More details can be found in Table 6 (Appendix A.1).

Evaluation Metrics Following previous work (Barhom et al., 2019; Held et al., 2021), we conduct a comprehensive comparison using metrics including MUC, B^3, CEAF$_e$, CoNLL, and LEA. The CoNLL F1 is a composite metric representing the average of the first three. B^3 is chosen for analysis, following Held et al. (2021).

Hyper Parameters For LLM summarization, we use the "`GPT-4-0613`" model via OpenAI API, setting the sampling temperature $t = 0$ to reduce the impact of randomness. In the first step of the generation workflow, we introduce a pre-step of instructing the LLM to perform dependency parsing on the sentence containing the event mention. Based on the parsing results, the LLM then elaborates on the mention. For SLM integration, we employ the pre-trained RoBERTa$_{\text{LARGE}}$ model (Liu et al., 2019) to embed event mentions, following our baseline (Held et al., 2021). For all three datasets, we apply a consistent set of hyper-parameters for fine-tuning, as detailed in Table 7 (Appendix A.2). In all experiments, be it primary results or analyses, we ensure reliability by conducting three independent experiments and averaging the outcomes.

4.2 Results

The main results are presented in Table 2. Our method achieves new state-of-the-art results on all three datasets, outperforming both the previously reported best results and the improved results obtained by our reproduced baseline.

ECB+ On this widely studied dataset, our method demonstrates improvements of 1.5% in CoNLL F1, compared to our baseline. In comparison to Held et al. (2021), upon which our baseline is built, our method also exhibits a 1% increase in CoNLL F1. [4] This improvement stands out notably in recent research, accompanied by significance testing to demonstrate its robustness. Compared to Chen et al. (2023), who also employs RoBERTa$_{\text{LARGE}}$ for encoding while proposing a different method to leverage broader contexts, we also achieve a 1% improvement, showcasing the effectiveness of our method in utilizing context.

GVC & FCC Our method demonstrates improvements of 2.7% and 7.0% in CoNLL F1 on the GVC and FCC datasets, respectively, compared to our baseline. The significant improvement on the challenging FCC dataset further underscores the effectiveness of our collaborative approach in leveraging LLM. Additionally, our baseline also shows improvements of 1.0% and 7.3% compared

[4] Our baseline is slight lower in CoNLL F1 than reported by Held et al. (2021), potentially be attributed to randomness.

Methods	MUC			B^3			$CEAF_e$			CoNLL	LEA		
	R	P	F1	R	P	F1	R	P	F1	F1	R	P	F1
ECB+													
Barhom et al. (2019)	77.6	84.5	80.9	76.1	85.1	80.3	81.0	73.8	77.3	79.5	-	-	-
Cattan et al. (2020)	85.1	81.9	83.5	82.1	82.7	82.4	75.2	78.9	77.0	81.0	-	-	-
Bugert et al. (2021)	76.0	76.1	76.1	71.8	81.2	76.2	72.2	72.1	72.2	74.8	55.1	67.9	60.8
Caciularu et al. (2021)	87.1	**89.2**	88.1	84.9	**87.9**	86.4	**83.3**	81.2	82.2	85.6	76.7	77.2	76.9
Held et al. (2021)	87.0	88.1	87.5	85.6	87.7	86.6	80.3	**85.8**	82.9	85.7	74.9	73.2	74.0
Hsu and Horwood (2022)	87.8	82.9	85.3	86.5	83.1	84.8	76.9	82.8	79.7	83.3	74.4	74.0	74.2
Yu et al. (2022)	88.1	85.1	86.6	86.1	84.7	85.4	79.6	83.1	81.3	84.4	-	-	-
Ahmed et al. (2023)	80.0	87.3	83.5	79.6	85.4	82.4	83.1	75.5	79.1	81.7	70.5	73.3	71.9
Chen et al. (2023)	88.6	85.9	87.2	87.8	85.4	86.6	82.8	83.7	83.2	85.7	-	-	-
Our baseline	86.6	86.8	86.7	87.1	86.0	86.5	82.6	82.5	82.5	85.2	77.8	76.6	77.2
Our method	**89.4**	87.1	**88.2**	**89.1**	86.5	**87.8**	82.7	85.5	**84.1**	**86.7**	**79.7**	**78.5**	**79.3**
GVC													
Barhom et al. (2019)	-	-	-	81.0	66.0	72.7	-	-	-	-	-	-	-
Bugert et al. (2021)	66.3	78.1	71.7	49.9	73.6	59.5	60.9	38.2	47.0	59.4	38.2	56.5	45.6
Held et al. (2021)	91.8	91.2	91.5	82.2	83.8	83.0	75.5	77.9	76.7	83.7	79.0	82.3	80.6
Ahmed et al. (2023)	84.0	91.1	87.4	79.0	76.4	77.7	69.6	52.5	59.9	75.0	74.1	63.9	68.6
Our baseline	91.3	92.0	91.7	86.2	83.8	84.9	78.7	76.5	77.6	84.7	82.0	78.4	80.2
Our method	**92.4**	**93.2**	**92.8**	**87.0**	**87.4**	**87.2**	**83.6**	**80.7**	**82.1**	**87.4**	**83.4**	**83.0**	**83.2**
FCC													
Barhom et al. (2019)	-	-	-	36.0	**83.0**	50.2	-	-	-	-	-	-	-
Bugert et al. (2021)	82.7	78.3	80.4	70.8	38.3	49.2	28.2	40.4	33.2	54.3	60.4	30.4	39.8
Held et al. (2021)	**86.4**	75.7	80.7	61.6	65.4	63.5	39.1	**65.3**	48.9	64.4	47.2	57.0	51.6
Our baseline	81.4	89.0	85.1	69.4	66.6	68.0	76.4	52.2	62.0	71.7	63.5	54.6	58.7
Our method	85.3	**90.6**	**87.8**	**74.5**	82.5	**78.3**	**80.9**	61.5	**69.8**	**78.7**	**69.7**	**73.5**	**71.5**

Table 2: Performance comparison on the ECB+, GVC, and FCC datasets. Our baseline results are obtained by replicating the state-of-the-art method proposed by Held et al. (2021), with the adoption of more advanced hyper-parameters. Our method shows a statistically significant improvement compared to our baseline, with a significance level of $p < 0.01$. The best results are highlighted in bold.

to Held et al. (2021), highlighting our comprehensive exploration on these two less-studied datasets.

4.3 The Impact of LLM Summarization

Error Analysis To gain a deeper understanding of the improvements achieved through LLM summarization, we perform a quantitative analysis on the false links within the clusters (see Table 3).

Similar to Yu et al. (2022), we categorize link errors into two types: false positive (FP) and false negative (FN). FP links (incorrect links) occur when two non-coreferential mentions are clustered together, while FN links (missing links) occur when two coreferential mentions are not clustered together. Additionally, we further categorize FP links into two sub-types based on whether two mentions share the same event type.[5] FPA (false positives caused by arguments) indicates that two mentions of the same type differ in argument information. FPT (false positives caused by types) implies that two mentions actually belong to different event types, eliminating the need to consider arguments.

[5]Details are provided in Appendix A.3.

Dataset	Method	FPA	FPT	FN
ECB+	Our baseline	1775	302	1262
	Our method	1227	152	1087
GVC	Our baseline	1412	13	1041
	Our method	865	13	1173
FCC	Our baseline	38522	0	8978
	Our method	4037	20	8575

Table 3: Statistics of errors by different types.

FPA Our method demonstrates the most substantial reduction in FPA errors across all three datasets, making the greatest contribution to the overall improvement. The reduction is approximately 30% for both ECB+ and GVC, and nearly 90% for FCC. The significant reduction on FCC is primary attributed to its nature, comprising multiple consecutive events from a large tournament, resulting in more pronounced contextual similarities. This underscores the effectiveness of our method in distinguishing events with similar contextual narratives (aligning with the challenge 1 from introduction). In Table 4, we present instances illustrating two highly similar earthquakes. The original context

Context	Summarization
Dozens injured, child dead as [*6.1 - magnitude earthquake*] hits Indonesia's Aceh Updated : July 02, 2013 15:50 IST A 6.1-magnitude earthquake which hit the Indonesian province of Aceh on Tuesday killed a child, injured dozens and destroyed buildings ... The quake struck inland at 0737 GMT at a depth of just 10 kilometres (6.2 miles) ... the US Geological Survey said . House collapsed ... 50 people with injuries ... 30 people seriously injured ... People panicked and rushed out of their homes ... In 2004 a massive tremor sparked a tsunami ...	[*6.1-magnitude earthquake*] refers to the seismic event that occurred in the **Bener Meriah district in the heart of Aceh**, Indonesia, on **July 2, 2013**. The earthquake struck inland at 0737 GMT at a depth of just 10 kilometres (6.2 miles) and was felt strongly for around 15 seconds, **from Bener Meriah to Banda Aceh**.
Indonesia's West Papua province was hit by a magnitude 6.1 [*earthquake*] today, the latest powerful tremor to shake the region where five people were killed and hundreds injured at the weekend when buildings were destroyed. The quake struck off the coast at 7:48 a.m. local time, 75 kilometers (50 miles) ... the U.S. Geological Survey said ... At least five people were killed, 250 others injured and more than 800 homes destroyed ... 14,000 people fled their homes ... temblor in 2004 caused a tsunami ...	[*earthquake*] refers to the magnitude 6.1 earthquake that hit Indonesia's **West Papua** province on an **unspecified date**. The earthquake struck off the coast at 7:48 a.m. local time, 75 kilometers (50 miles) west of the region's **main city of Manokwari**, according to the U.S. Geological Survey.

Table 4: Two non-coreferential mentions for the event type "earthquake", illustrating the remarkably similar contexts, as well as our generated more distinctive summaries. To better illustrate the similarity, we preserve the sentence containing the mention along with similar content from the context. Key information in our summarization is highlighted in bold. Mention spans are represented as [*mention span*].

includes details about the earthquake occurrence, earthquake casualties, media coverage, and historical events. Our generated summaries primarily focus on the core details of the earthquakes, such as date and specific location, thus facilitating their differentiation. It can be observed that our LLM summarization is capable of identifying specific information for particular events and aggregating sufficient details from the entire context.

FPT Compared to FPA, there are significantly fewer FPT errors, only appearing in the ECB+ dataset. The few occurrences on the single-topic GVC and FCC datasets can likely be disregarded, possibly due to random factors. This is because in the multi-topic ECB+ dataset, there may be topics in the test set that were not encountered in the training set, leading to unseen event types. By reducing half of the FPT errors on the ECB+ dataset, it signifies that our summarization also assists in distinguishing unseen event types.

FN Our method shows less improvement in reducing FN errors compared to FP. The challenges arise from two primary factors. Firstly, mentions of the same event can vary greatly in expression styles. Secondly, some event mentions naturally lack sufficient details as the authors assume that readers already possess necessary background information. We illustrate these issues with instances in Appendix A.4. For these cases (aligning with the challenge 2 in introduction), additional training data or external information retrieval may be necessary, as our faithful summarization based on the original context struggles to cope.

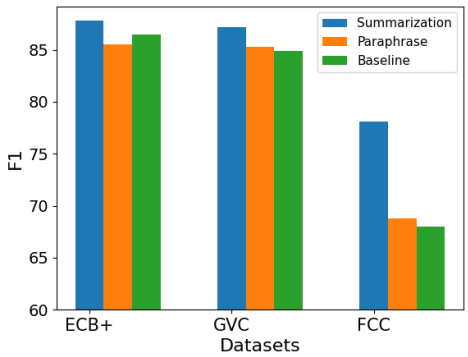


Figure 2: LLM paraphrase comparison with B^3 F1. The vertical axis has a baseline starting from 60.

Overall, LLM summarization excels in consolidating information for specific events, facilitating the differentiation of similar yet non-coreferential events. Relatively, its effectiveness is limited for events with significant expression differences or those lacking essential details.

LLM Summarization VS LLM Paraphrase To validate that the performance improvement brought by our summarization is due to genuinely extracting crucial information rather than introducing diversity in context, we conduct a comparison with paraphrases generated by the LLM. We prompt the LLM to paraphrase the context of mentions instead of the sentences they belong to, and use the same hyper-parameters for fine-tuning the SLM.

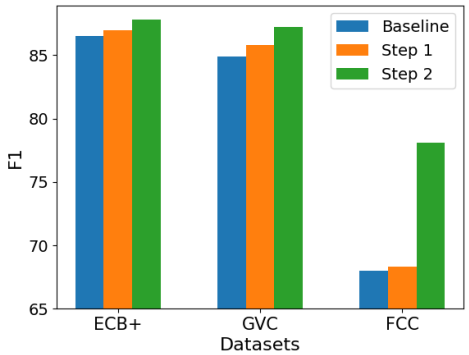


Figure 3: Comparison of different steps with B^3 F1. The vertical axis has a baseline starting from 65.

As shown in Figure 2, compared to our baseline, LLM paraphrase exhibits a slight improvement on GVC and FCC, with a more pronounced decline on ECB+. More importantly, it significantly lags behind our summarization on all datasets. This comparison demonstrates the capability of our summarization method in selecting and aggregating relevant information. The prompt for LLM paraphrase is provided in Table 11 (Appendix A.5).

4.4 Ablation Study on the Two-step Workflow

We conduct an ablation study to specifically illustrate the effect of Step 1 and Step 2 in LLM summarization (Table 1). As shown in Figure 3, both steps contribute to the overall improvement, with the second step being more pronounced, especially on the FCC dataset. This is attributed to the longer documents in FCC, with nearly double the number of sentences in each document compared to the other two datasets. This demonstrates that the information provided in Step 1 establishes a solid foundation but is relatively localized. Step 2, involving global information expansion, plays a crucial role in overall enhancement.

To examine the benefits of decomposed execution, we further integrate the two-step workflow into a single-step one through simple concatenation. Despite demonstrating comparable performance on GVC, the integrated workflow shows a noticeable lag, being 1.2% and 2% behind on ECB+ and FCC, respectively, in terms of B^3 F1. This indicates that even with straightforward instructions, decomposing the multi-objective task into multiple independent steps is necessary, as evidenced by the recent LLM agent studies (Aksitov et al., 2023).

We perform error analysis and compare the lengths of the generated summaries to provide a detailed explanation of the impact of each step in the workflow and its decomposition. Further details can be found in Appendix A.6.

Method		R	P	F1
Cluster+Lemma (Barhom et al., 2019)		71.7	85.0	77.8
Few-shot	Mention-inclusive sentences	76.3	78.1	77.2
	Full context	65.6	77.2	70.9
Zero-shot	Mention-inclusive sentences	78.6	60.4	68.3
	Full context	75.6	56.4	64.6

Table 5: Results on the ECB+ test set, based on the B^3 metric. "Mention-inclusive sentences" indicates that we retain only those sentences containing mentions, reducing the complexity of contextual understanding.

4.5 GPT-4 Performance on CDECR

We test the performance of GPT-4 using in-context learning in both few-shot and zero-shot settings (see Table 5). In the few-shot setting, GPT-4 achieves performance comparable to the method based on lemma matching (Here lemma refers to the span words of a mention.). According to our observations, GPT-4 still relies on a simplistic approach of clustering based on the literal meaning of mentions. And the role of demonstrations seems limited to expanding the scope of matching for synonymous mentions without considering their contexts. Moreover, its performance notably drops when dealing with complete contexts, showing difficulty in understanding lengthy contexts composed of multiple documents. In the zero-shot setting, there are cases where a large number of unrelated mentions are clustered into a single category, resulting in relatively high recall but very low precision. This emphasizes the complexity of the CDECR task, as GPT-4 faces challenges in performing basic clustering solely relying on task description. Implementation details can be found in Appendix A.7.

5 Conclusion

We design generic tasks to leverage the potential of LLMs for CDECR, effectively bridging the gap between the general capabilities of LLMs and the complex annotation guidelines of specific IE tasks. Results show that by harnessing the inherent knowledge and comprehension abilities of LLMs to gain a deeper understanding of events, our collaborative approach can alleviate the challenge of SLMs for complex contextual understanding, ultimately enhancing performance.

Limitations

The LLM we use is `GPT-4-0613`. Moving forward, we plan to assess the performance of additional LLMs, such as LLaMa (Touvron et al., 2023).

For CDECR, where internal information within the given document might be insufficient, there arises a need for external information retrieval. We are considering further leveraging the capabilities of LLMs to explore how to retrieve supplementary information from external corpora such as news articles. Our aim is to combine this additional information with the given documents to enhance performance.

Ethics Statement

We adhere to the ACL Code of Ethics.

References

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. $2 * n$ is better than n^2: Decomposing event coreference resolution into two tractable problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1569–1583, Toronto, Canada. Association for Computational Linguistics.

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, et al. 2023. Rest meets react: Self-improvement for multi-step reasoning llm agent. *arXiv preprint arXiv:2312.10003*.

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4659–4671, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *arXiv preprint arXiv:2009.11032*.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. Association for Computational Linguistics.

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4833–4843, Singapore. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

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

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Hsu and Graham Horwood. 2022. Contrastive representation learning for cross-document coreference resolution of events and entities. In *Proceedings*

of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3644–3655, Seattle, United States. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Samuel Kriman and Heng Ji. 2021. Joint detection and coreference resolution of entities and events with document-level context aggregation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 174–179, Online. Association for Computational Linguistics.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, Jeju Island, Korea. 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. *arXiv preprint arXiv:2304.11633*.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023b. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore. Association for Computational Linguistics.

Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Takao Osaki, Katsushi Matsuda, Haifeng Chen, et al. 2023. Improving open information extraction with large language models: A study on demonstration uncertainty. *arXiv preprint arXiv:2309.03433*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: a survey of two decades of research. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5479–5486.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. Association for Computational Linguistics.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4840–4850, Online. Association for Computational Linguistics.

Ji Qi, Chuchun Zhang, Xiaozhi Wang, Kaisheng Zeng, Jifan Yu, Jinxin Liu, Lei Hou, Juanzi Li, and Xu Bin. 2023. Preserving knowledge invariance: Rethinking robustness evaluation of open information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Singapore. Association for Computational Linguistics.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1708–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Piek Vossen and Agata Cybulska. 2018. Identity and granularity of events in text. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, pages 501–522. Springer.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: a data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

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

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2022. Improving event coreference resolution using document-level and topic-level information. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6765–6775, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 190–200, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 69–78, Seattle, Washington. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

A Experimental details

A.1 Dataset Statistics

As shown in Table 6.

	ECB+	GVC	FCC
Documents	982	510	451
Sentences	16314	9782	14940
Event mentions	6833	7298	3563
Event clusters	2741	1411	469
Event coref links	26712	29398	145272

Table 6: Statistics of each dataset.

A.2 SLM Fine-tuning Hyper-parameters

As shown in Table 7.

	Candidate retrieval	Pairwise classification
Learning rate	1e-5	6e-6
Batch size	16	16
Epochs	50	20
Early stop patience	10	5
Train neighbor size	-	20
Eval neighbor size	10	10

Table 7: Hyper-parameters for fine-tuning the SLM-based modules.

A.3 Event Type Categorization

	Mentions	Clusters	Types
ECB+	1780	805	405
GVC	1008	194	4
FCC	1074	167	19

Table 8: Statistics for mention, cluster, and event type in the test set.

To categorize event types, we establish a three-layer hierarchical structure of (mention->cluster->type), linking types between mentions. Specifically, if there are synonymous mentions between any two clusters, they belong to the same event type, and all mentions within the clusters belong to a synonymous event type. Drawing inspiration from (Ahmed et al., 2023), we determine mention synonymity by matching their span words. Table 8 illustrates that the contents of FCC and GVC belong to the same topic, resulting in a concentrated set of event types. Conversely, ECB+ involves various topics such as quake, murder, acquisition, etc., leading to a diverse set of event types.

A.4 False Negative Cases

Given the context where mentions of the same event can vary greatly in expression styles, we provide an illustrative example in Table 9.

Event	Smith case as the incarnation of the Doctor
Mention expressions	was handed the keys to the Tardis winning the role of the 11th Doctor stepping into Doctor Who's title role

Table 9: Variations in mention expressions for identical event.

In cases where event mentions naturally lack sufficient details, we illustrate this phenomenon through Table 10, which presents two mentions of the same earthquake. The context for the first mention contains essential information such as time, location, magnitude, casualties, etc. In contrast, the context for the second mention primarily describes the subjective experiences of the individuals involved, lacking details related to the event itself. Despite our summarization extracting key information from the original context, it encounters difficulties in supporting coreference judgments.

A.5 LLM Paraphrase Prompt

As shown in Table 11.

A.6 Two-step Workflow Analysis

Error Analysis We conduct error analysis for the workflow with only Step 1, the complete two-step workflow (Step 2), and the integrated single-step workflow.

As shown in Figure 4, Step 1 exhibits a significant reduction in FPA errors across all three datasets, indicating its effectiveness in extracting tailored information. However, an increase in FN errors is observed across all three datasets, suggesting that while Step 1 provides sufficiently distinctive information, it lacks the details needed to link mentions of the same event. This issue was notably addressed by the introduction of Step 2, resulting in a substantial decrease in FN errors across all datasets. FPA errors are also largely maintained at the level achieved in Step 1, leading to a significant improvement in coreference results. This emphasizes the indispensable roles of both Step 1 and Step 2 in the final outcomes. In Table 12, we provide examples to compare summaries generated by Step 1 and Step 2.

Compared to the two-step workflow, the integrated single-step workflow shows differing degrees of increase in both FPA and FN errors, further underscoring the necessity of decomposed execution.

Summarization length comparison We further compare the lengths of summaries generated in Step 1 and Step 2. As illustrated by the green line in Figure 5, it is evident that Step 2, building upon Step 1, results in approximately double the length. The additional detailed content contributes to the reduction of FN errors, effectively linking mentions of the same event. Furthermore, as indicated by the red and blue lines, our generated summaries remain within approximately 20% of the original document starting from a document length of 200 words. Moreover, with the increase in document length, this proportion further diminishes. This reflects the conciseness our summarization.

A.7 GPT-4 Evaluation

Considering the need to handle very long contexts composed of multiple documents, we opt for the "`GPT-4-Turbo-Preview`" model from OpenAI, which can accommodate up to 120k tokens. Despite this, the model still cannot accommodate all 200+ documents in the ECB+ test set. Hence, we conduct testing at the topic level, which means placing all documents from a topic into a single prompt (there are no cross-topic links in ECB+). Also, We introduce only one random selected topic from the training data as the demonstration. The prompt is shown in Table 13. We set the model parameters, including seed and temperature, to 0 to minimize randomness. Additionally, we specify the output format to be in JSON.

Context	Summarization
[*4.6 earthquake*] shakes Northern California March 14, 2013 | 8:16 am A magnitude 4.6 earthquake struck Northern California on Thursday morning, The temblor struck about 26 miles north of Santa Rosa in the Geysers area. There were no reports of damage or injuries, but there were some apparent aftershocks. It was felt over a wide area of the region, including in Santa Rosa, Marin County and Vallejo, according to the U.S. Geological Survey. The "Do You Feel It?" survey by the USGS showed people as far south a San Francisco felt shaking.	[*4.6 earthquake*] refers to the seismic event that occurred in **Northern California** on **March 14, 2013**, with a magnitude of 4.6 on the Richter scale. The earthquake was felt over a wide area of the region, including in **Santa Rosa, Marin County, and Vallejo**, and caused some apparent aftershocks. However, there were no reports of damage or injuries.
Good sized quake at the California Geysers I felt this one pretty good. I'm located 25 miles south of the quake. Slow easy rocking for about 6 seconds. This one is pretty rare for the Geysers. The Geysers is one of the most active areas of Northern California with small earthquakes hitting the area nearly every week. A 4.4 [*quake*] is on the large size for The Geysers. The UGSG Website says the largest quake ever to be recorded there was 4.5.	[*quake*] refers to a seismic event that occurred at the **California Geysers**, which was felt **25 miles south of the epicenter**. The quake was described as a slow, easy rocking that lasted for about 6 seconds. The **exact date of the quake is not provided** in the news, but it can be inferred to have occurred recently based on the publication date of the news.

Table 10: Two coreferential mentions referring to the same earthquake, where the second provides minimal coreference evidence. Key information in our summarization is highlighted in bold. Mention spans are represented as [*mention span*].

News: `[input document]`
Question 1: In this news, given "`[mention 1]`" mentioned in the sentence "`[the sentence]`".
Concatenate the preceding five sentences of the current sentence (ignore if not available), the current sentence, and the subsequent five sentences of the current sentence (ignore if not available) into a single paragraph. Then, paraphrase the concatenated paragraph while preserving the mention `[mention 1]`. Attempt to express the information differently while maintaining the meaning and key information. Ensure that the mention `[mention 1]` is preserved and marked as `#[mention 1]#` in the paraphrased result. Limit the paraphrased result to three sentences. Present the information in the following format: 'Paraphrase: <placeholder>'.
Question 2: ...

Table 11: The prompt for LLM paraphrase. Each prompt includes a document along with multiple event mentions. The content to be filled is represented as `[content]`.

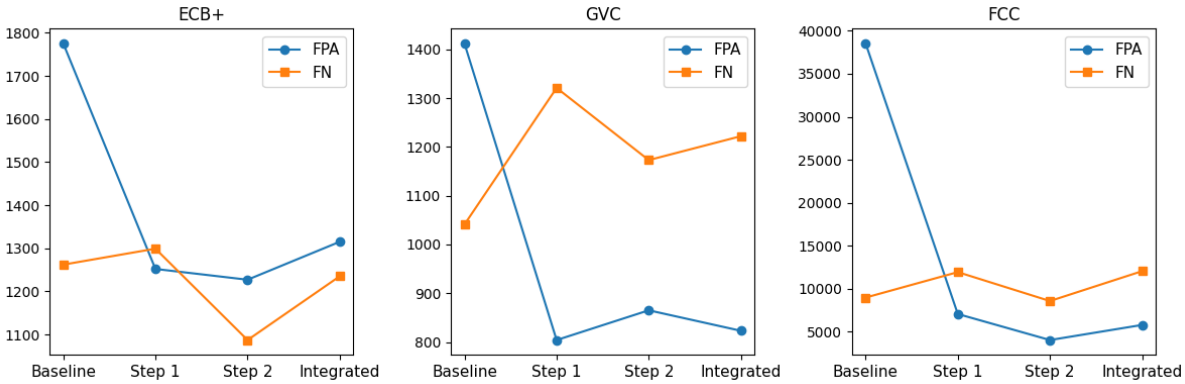


Figure 4: FPA and FN error comparison. Due to the rarity of FPT-type errors, we have omitted them in the figures for better clarity in presentation. Step 2 is built upon Step 1, and the integrated involves merging the two steps together.

Step 1	Step 2
[*6.1-magnitude earthquake*] refers to the seismic event that occurred in Aceh, Indonesia, with a magnitude of 6.1 on the Richter scale.	[*6.1-magnitude earthquake*] refers to the seismic event that occurred in the **Bener Meriah district in the heart of Aceh**, Indonesia, on **July 2, 2013**. The earthquake struck inland at 0737 GMT at a depth of just 10 kilometres (6.2 miles) and was felt strongly for around 15 seconds, **from Bener Meriah to Banda Aceh**.
[*earthquake*] refers to the magnitude 6.1 earthquake that hit Indonesia's West Papua province.	[*earthquake*] refers to the magnitude 6.1 earthquake that hit Indonesia's **West Papua** province on an **unspecified date**. The earthquake struck off the coast at 7:48 a.m. local time, 75 kilometers (50 miles) west of the region's **main city of Manokwari**, according to the U.S. Geological Survey.

Table 12: Comparison of summaries generated by Step 1 and Step 2. Step 2 is built upon Step 1. Key information for distinguishing in Step 2 is highlighted in bold. Mention spans are represented as [*mention span*].

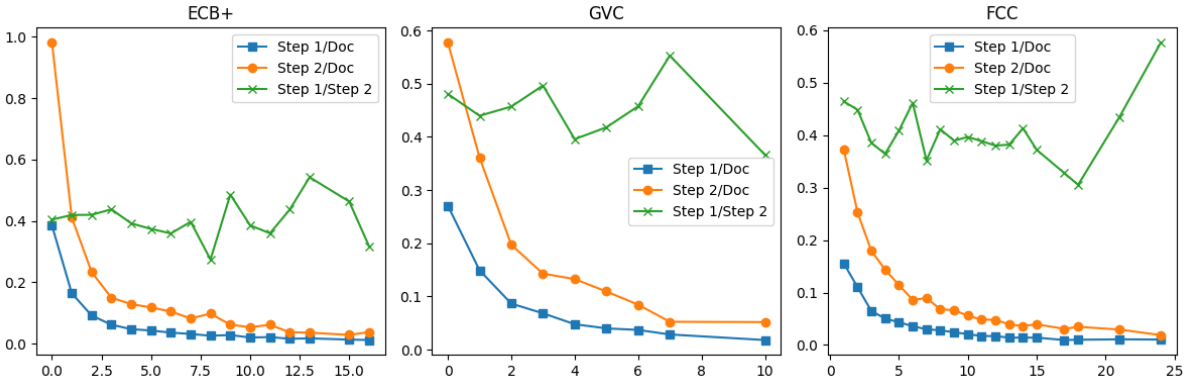


Figure 5: Summarization length comparison. Step 2 is built upon Step 1. The vertical axis represents the ratio of content word count. The horizontal axis represents the number of words in the content, scaled by a factor of 100.

System role	You are a helpful assistant tasked with clustering coreferential event mentions in the provided documents. The event mentions in the documents are marked as follows: [mention string](mention id). Please output the result in JSON format without whitespace. In the JSON structure, each 'mention id' is assigned a 'cluster id'.
Prompt	You can learn from the following example: Input: Document: `[... [mention](mention_id) ...]` ... Output: `[mention_id: cluster_id, ...]` Now the following is your task: Document: `[... [mention](mention_id) ...]` ...

Table 13: The prompt for GPT-4 evaluation. The system role is used to declare task requirements and output specifications. The prompt is divided into two sections: initially, a demonstration, followed by data to be processed.