# Cross-Document Event Coreference Resolution: Instruct Humans or Instruct GPT?

**Jin Zhao**
Brandeis University
jinzhao@brandeis.edu

**Nianwen Xue**
Brandeis University
xuen@brandeis.edu

**Bonan Min**[*]
Amazon AWS AI Labs
bonanmin@amazon.com

## Abstract

This paper explores utilizing Large Language Models (LLMs) to perform Cross-Document Event Coreference Resolution (CDEC) annotations and evaluates how they fare against human annotators with different levels of training. Specifically, we formulate CDEC as a multi-class classification problem on pairs of events that are represented as decontextualized sentences, and compare the predictions of GPT-4 with the judgment of fully trained annotators and crowdworkers on the same dataset. Our study indicates that GPT-4 with zero-shot learning outperformed crowd-workers by a large margin and exhibits a level of performance comparable to trained annotators. Upon closer analysis, GPT-4 also exhibits tendencies of being overly confident, and forcing annotation decisions even when such decisions are not warranted due to insufficient information. Our results have implications on how to perform complicated annotations such as CDEC in the age of LLMs, and show that the best way to acquire such annotations might be to combine the strengths of LLMs and trained human annotators in the annotation process, and using untrained or undertrained crowdworkers is no longer a viable option to acquire high-quality data to advance the state of the art for such problems. We make our source and data publicly available.[1]

## 1 Introduction

Cross-Document Event Coreference Resolution (CDEC) is the task of identifying coreferent events from different documents. Traditionally, CDEC has been approached as a mention-pair classification problem, in which the goal is to determine if two event mentions refer to the same event based on their contextual information in their containing articles (Lu and Lu, 2021), followed by a clustering step where coreferent events are grouped into clusters. CDEC is a challenging problem for both data annotation and computational modeling because, in order to determine whether two event mentions are coreferent with each other, their participants, time, and location would have to be the same or at least compatible, and such information would have to be found in the global context of the document or even outside of the document. Another challenge for CDEC annotation is that in a random collection of documents, coreferent event pairs are likely to be very sparse. In order to make a CDEC annotation task feasible, it is necessary to pre-select the documents by their topics to increase the density of coreferring events, and to limit the length of the document to reduce the size of the context that needs to be examined. As a result, existing datasets such as ECB+ (Cybulska and Vossen, 2014), a widely used benchmark for CDEC, consists of relatively short articles and has limited coverage and diversity of event types.

In this paper, we explore CDEC data creation with regular sized news articles. We frame CDEC as a multi-class classification problem on pairs of events represented as sentences containing event trigger words. To make the task feasible, we perform decontextualizaton on these sentences to make them self-contained with the necessary contextual information. The decontextualized event sentence pair are illustrated in Figure 1. We also pre-select candidate event pairs using a state-of-the-art CDEC model CDLM (Caciularu et al., 2021) by maximizing the recall so that we don't unintentionally rule out coreferring event pairs. Performing these steps allows us to present candidate event pairs to annotators who can determine if they are coreferent based on just the event pairs. Given the impressive performance of GPT-based LLMs, several recent studies explored using GPTs to create annotated datasets for text generation (Wang et al., 2021) and social computing (Zhu et al., 2023). To investigate how well GPT-4 (OpenAI, 2023)

---

[*] Work done prior to joining AWS AI.

[1] https://github.com/jinzhao3611/CDEC

*Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 561–574
December 6–7, 2023. ©2023 Association for Computational Linguistics

**DOC 1**: [...] Despite the <departure> of protesters, it is not expected that the construction will resume in the near future. [...]
**Decontextualized event sentence**: Despite the <departure> of protesters from Mauna Kea Mountain, it is not expected that the construction of the Thirty Meter Telescope project will resume in the near future.
**DOC 2**: [...] In March, protesters who had been camped out on Mauna Kea to block construction of Thirty Meter Telescope dismantled their large tents and <left> because of concerns about the spread of the coronavirus. [...]
**Decontextualized event sentence**: In March, protesters, who who had been camped out on Mauna Kea mountain to block construction of Thirty Meter Telescope dismantled their large tents and <left> because of concerns about the spread of the coronavirus.

Figure 1: A CDEC example formulated as **decontextualized** event mention pair classification. Event trigger words are in angle brackets and the underlined text represents the inserted contextual information. Event represented by <departure> in DOC 1, and event represented by <left> in DOC 2 both refer to leaving actions taken by the protesters camped on Mauna Kea Mountain due to the spread of the coronavirus.

performs against human annotators with different levels of training, we gave the same data set of event mention pairs to fully trained annotators, crowd-workers, and GPT-4, requiring them to provide nine-way decisions. To establish our ground truth, we conducted adjudication meetings with our trained annotators to resolve disagreements. Subsequently, we calculated the accuracy of both human annotations and GPT-4 annotations against the ground truth.

Our findings reveal that GPT-4 outperformed crowd workers by a large margin and achieved a level of performance comparable to trained human annotators. Upon closer examination, our results show that for both human annotators and GPT-4, performance varies across topics and coreference categories. They also show that GPT-4 exhibits a strong tendency to make inferences even when there is not sufficient contextual basis, and bridge gaps in understanding by resorting to hallucination. We believe that our work has implications for the creation of a complex and labor-intensive annotations such as CDEC. The far superior performance of GPT-4 against untrained crowd workers means that there is little value in performing the CDEC annotation in this setting. GPT-4 also has the potential to accelerate the annotation process by reducing the need for trained human annotators, resulting in significant time and cost savings. The most effective data creation process for complicated datasets in the age of LLMs might be one that combines strengths of LLMs and trained annotators. We leave it to future research exactly how that should be carried out.

The rest of the paper is organized as follows. In §2, we discuss related work. In §3, we describe our data preparation process. We present and discuss our experimental results in §4, and conclude in §5.

## 2 Related Work

### 2.1 Current practice in CDEC Data creation

A number of CDEC datasets have recently been created and they include MEANTIME (Minard et al., 2016), EER (Hong et al., 2016), and RED (O'Gorman et al., 2016). When annotating such datasets, annotators must exhaustively compare each event mention in the dataset against all other event mentions across documents to establish coreference relations. This is a labor-intensive process and as a result, existing datasets are all relatively small. By representing events as decontextualized sentences that can stand alone, there is the potential to create CDEC datasets on a much larger scale, as annotators only need to examine a pair of sentences to make coreference decisions.

As co-referring events in text are often sparsely distributed, to make the annotation process feasible, it is often necessary to limit such annotation to certain topics and a small number of event types. For example, in ECB+ (Cybulska and Vossen, 2014), a widely used benchmark data set for CDEC, each topic focuses on a limited number of specific events, resulting in low variation in unique trigger words within each coreference cluster (averaged 2.66 unique trigger words per cluster). Although we also limit the topics in our CDEC annotation setup, there is no such limitation on the event types as we do not use a pre-defined list of trigger words to identify potentially coreferring candidate event pairs.

A lot of efforts are put in circumventing the scalability issue of manually created data by creating auto or semi-automatically annotated CDEC datasets. GVC (Vossen et al., 2018) marks event references using a structured database of known gun violence events in a semi-automatic fashion. It considerably improves annotation efficiency and event variation compared to ECB+, but it does not apply to broader data topics other than gun violence. HyperCoref (Bugert and Gurevych, 2021) and WEC-Eng (Eirew et al., 2021) leveraged article hyperlinks in Wikipedia data to create data automatically. However, there is no guarantee that the events marked by the Wikipedia contributors will be consistent. Moreover, they mainly consist of Wikipedia-entry worthy or what Eirew et al. (2021) call referential event mentions, but do not cover descriptive or anecdotal events that arise in news reports.

### 2.2 Annotations by GPTs

There have also been previous efforts in leveraging LLMs to obtain annotated data. In a study by Wang et al. (2021), it was reported that the use of GPT-3 generated labels for the annotation of data can lead to cost savings ranging from 50% to 96%, while maintaining comparable performance in various NLP tasks. Zhu et al. (2023) shows ChatGPT obtains an average accuracy 60.9% in multiple social computing tasks. Bang et al. (2023) conducted a comprehensive evaluation of ChatGPT, demonstrating its superior performance on various NLP tasks over other LLMs while highlighting its potential hallucination issues in reasoning tasks. Huang et al. (2023) examined the quality of ChatGPT-generated natural language explanations for implicit hateful speech, demonstrating that ChatGPT correctly identifies 80% of such tweets and suggesting its potential as a data annotation tool. However, none of these works attempted to use LLMs to annotate the data for CDEC.

## 3 Data Preparation

A number of critical steps need to be taken to prepare the data for annotation by human annotators and GPT-4, and they include source article selection and deduplication, event extraction, decontextualization, and pre-selection of candidate event pairs. A flow chart illustration of this pipeline can be found in Figure 4 in the Appendix A.3.

### 3.1 Data Sourcing

The data used in our study were obtained from AylienAPI[2], a platform that offers access to a vast Coronavirus dataset that contains more than 1,500,000 news articles related to the pandemic, starting in November 2019. To ensure the relevance and coherence of our dataset, we utilize AylienAPI's keyword feature to collect articles on specific topics of interest. We manually select 100 articles across 10 different topics from this dataset, aiming to include articles from diverse news sources to enrich variety of trigger words within our dataset.

### 3.2 Data Preprocessing

To address the substantial duplication found in aggregated news articles, we employ LSH (locality sensitive hashing)[3] for document deduplication. This process effectively identify and remove duplicate documents, reducing redundancy within our dataset. Additionally, we exclude editorials that express subjective opinions on topical issues and eliminate articles that provided briefings consisting of a collection of short news items. To further refine the dataset, we utilize regex-based filtering to exclude irrelevant events, specifically filtering out noise sentences like *comment below if you have any questions*. These steps are crucial in ensuring that the final dataset is of high quality and meet our research objectives.

### 3.3 Event Extraction

CDEC deals with identifying and clustering together textual mentions across multiple documents that refer to the same event. They include *descriptive event mentions*, which are typically expressed through verbs or nominalizations (e.g., “contracted the virus”, “analysis”) to provide new information, and *referential event mentions*, which are usually represented by noun phrases (e.g., “earthquake”, “Blizzcon 2019”) (Eirew et al., 2021) to provide a point of reference. We extract both types of event mentions from the dataset using the event extraction model proposed in Yao et al. (2021).

### 3.4 Decontextualization: Making the Events Stand Alone

CDEC often requires understanding the event that a sentence represents within a broader context, as crucial details such as participants, time, location, etc., might not be explicitly mentioned in a local textual window. Including entire documents that contain the candidate event mention pair can be costly when they are too long for both annotation tasks or computational modeling. To address this, we employ event decontextualization, a technique that renders events interpretable even when taken out of the document context, while preserving their intended meaning. We utilize the fine-tuned T5 (Raffel et al., 2020) model described in Choi et al. (2021) to perform decontextualization on sentences

---

[2] `https://aylien.com/`

[3] `https://github.com/kayzhu/LSHash`

containing event triggers. This model decontextualizes sentences by incorporating relevant context information from the document context. In the following example, we can decontextualize the original sentence by replacing "She" with "Dr. Calderwood", adding location or context "in a statement released on Sunday", where the mentioned information was provided.

> **Original**: She also said she would work to ensure a smooth transition to her successor.
> **Decontexualized**: *Dr. Calderwood* also said she would work to ensure a smooth transition to her successor *in a statement released on Sunday*.

### 3.5 Event Pairs Pre-selection

We use the CDLM (Caciularu et al., 2021), a pretrained cross-document language model, to select candidate event sentence pairs. CDLM incorporates the learning of cross-document relationships and utilizes dynamic global attention to predict masked tokens. In our experiments, we utilize CDLM for event pairwise scoring and pre-selection of the top-ranked event mention pairs. We select the top 200 event mention pairs from each of the 10 topics, resulting in a total of 2,000 pairs.

## 4 Experiments

### 4.1 Setup

We approach CDEC as a nine-class classification problem as shown in Table 1: "Identity", "Concept-Instance", "Instance-Concept", "Set-Member", "Member-Set", "Whole-Subevent", "Subevent-Whole", "Not-Related", "Cannot-Decide". Other than "Identity", "Not-Related", and "Cannot-Decide", the rest of them are symmetrical relations. Specifically, we have incorporated the "Identity", "Whole-Subevent", and "Set-Member" relations from the RED framework (O'Gorman et al., 2016) and the "Concept-Instance" relation from the confirmation relation in EER (Hong et al., 2016)

In order to achieve better agreement among annotators, we intentionally instruct annotators to disregard tense, aspect, and modality when making annotation decisions. For instance, annotators are specifically guided to annotate coreference between statements such as "Boris Johnson said he would <shake> hands with corona patients during that hospital visit on March 3" and "Boris Johnson <shook> hands with corona patients during that hospital visit on March 3". Although these statements do not strictly refer to the same event, they represent interesting event relations and can be filtered out using modality detection tools if they are deemed not to be true cases of event coreference.

#### 4.1.1 Trained Annotators

Four trained annotators, who are computational linguistics graduate students with prior experience in working with events, were hired in the annotation process. They underwent a comprehensive training process consisting of one hour of guideline training, a practice batch, and an adjudication meeting to resolve any discrepancies before proceeding with the actual annotation. The annotation guidelines can be found in Appendix A.2. During the annotation process, each annotator is assigned to work on one batch at a time. Each batch requires three annotators. Annotators are instructed to assume that all pairs of sentences within their assigned batch referred to the same microworld related to the given topic (Vossen et al., 2018). After completing each batch, an adjudication meeting was conducted to address any remaining differences and ensure consistency in the future annotations.

#### 4.1.2 Crowd Workers

For our crowdsourcing experiment, we utilize the Amazon Mechanical Turk platform[4]. We develop an interface that catered to both Turkers and trained annotators. Turkers are required to read the annotation guidelines and annotate the event pairs batch by batch, with each batch consisting of 200 event pairs. The Turkers are asked to choose from the same nine options, and each batch is assigned to 3 Turkers. Taking into account the complexity, time required, and market rates, we paid $0.1 per question in the screening stage, and $0.2 per question in the annotation stage.

In the screening stage, we publish a set of data already adjudicated. After rounds of monitoring Turkers' progress, providing feedback to guide their work, and initiating regular communications to address any questions or concerns they may have, we eventually selected 6 out of 56 Turkers who achieve at least 80% in accuracy in the "Not-Relate" category. This category was chosen as it requires more attention to get it right and it allows us to filter out potentially malicious Turkers. We subsequently contact these Turkers, provide them with feedback on their annotations, and invite them to work on additional batches following a similar process as our trained annotators. Our screening and

---

[4] `https://www.mturk.com/`

| Nine-Category | Explanation | Example |
|---|---|---|
| Identity<br>(O'Gorman et al., 2016) | Two event mentions refer to the same event | 1. But reduced punishment of gamer Blitzchung didn't stop angry Blizzard fans , who saw the initial <move> as overreach and a sign the Blizzard company had turned on them.<br>2. Following the gamer blitzchung <ban> by Blizzard, gamers Wright , Chambers and their third teammate , Corwin Dark , held a sign up on a collegiate Hearthstone livestream . |
| Concept-Instance/Instance-Concept<br>(Hong et al., 2016) | One event mention represents a generalized concept<br>The other is an concrete instance of the previous one | 1. On Oct. 8 , Blizzard <banned> Hearthstone pro Chung B̈litzchung N̈g Wai after he expressed support for Hong Kong protesters focused on democratic rights.<br>2. That <punishment> was shortened to a six-month suspension and gamer blitzchung 's prize money was honored after online outrage. |
| Set-Member<br>(O'Gorman et al., 2016) | One event mention represents a collection of events<br>The other is a subset or a member of previous one | 1. CNET sister site GameSpot will be covering the protests at Blizzcon, as well as each of the <announcements> from the show throughout the weekend .<br>2. Blizzard president Brack's <announcement> may not calm some fans , who 've been organizing online and planning to protest during the blizzcon event . |
| Whole-Subevent<br>(O'Gorman et al., 2016) | One event mention represent a larger event<br>The other is a component of the previous one | 1. Now , BlizzCon , the highly anticipated annual <convention> run by the company 's Blizzard Entertainment division , may be disrupted by demonstrations.<br>2. The opening <ceremony> of Blizzcon is usually streamed live on Blizzcon 's website. |
| Not-Related | Two event mentions are not related in any way above | 1. Those two students , Torin Wright and Casey Chambers , were the center of attraction for the protest at Blizzcon and gave individual speeches that were <met> with loud applause .<br>2. Blizzard president J. Allen Brack's statement was <met> with a round of applause from the Blizzcon audience . |
| Cannot-Decide | Cannot decide due to lack of information | 1. Kim did not publicly comment on the controversy and has continued to <restock> the collection of facial masks , although it is currently sold out yet again .<br>2. Claiming that as soon as the brand <restocked> Kim would donate the generous sum to those affected by the global pandemic. |

Table 1: CDEC as a Multi-Class Classification Task.

training methods for Turkers were based on the approaches outlined in Pyatkin et al. (2020) and Roit et al. (2019).

#### 4.1.3 GPT-4

In our experiments, we employ GPT-4, the latest model in the GPT series. We conduct zero-shot experiments with the `gpt-4` model using OpenAI API[5]. We provided GPT-4 prompts like the example prompt in Table 2.

Similarly, the trained annotators and Turkers are given the same set of questions and answer choices as GPT-4. Additionally, guidelines are provided to offer detailed explanations and examples for each answer choice to ensure consistent and accurate annotations, as illustrated in Figure 2 and Figure 3 in appendices.

What is the relation between the two marked events in the following sentence pair:
1. Isabel Dos Santos has since <left> Angola — along with several other members of the family — because she claims she has faced death threats.
2. The order said the central bank would ensure that no funds <leave> the personal bank accounts of the three accused.
The relation has to be one of the following: Identity, Concept-Instance, Instance-Concept, Set-Member, Member-Set, Subevent-Whole, Whole-Subevent, Not-Related, Cannot-Decide.
Provide an explanation

Table 2: An example prompt provided to GPT-4.

### 4.2 Evaluation

For the evaluation of our annotations, we calculate the Fleiss' Kappa (Fleiss, 1971) and WAWA score (Ning et al., 2018) for trained annotators. Fleiss' Kappa is a chance-corrected measure that assesses the level of agreement among more than two annotators. The WAWA score measures the agreement between each annotator with the ma-

---

[5] `https://platform.openai.com/docs`

| Method | Accuracy | IAA | |
|---|---|---|---|
| | | Fleiss' Kappa | WAWA |
| Trained Annotator | 69.85 | 48.79 | 74.40 |
| GPT-4 Zero-Shot | 64.00 | N/A | N/A |
| Turker | 42.65 | N/A | 52.50 |

Table 3: Accuracy Scores: IAA metric used for Trained Annotators is Fleiss' Kappa, and IAA metric used for Turkers is WAWA

jority consensus, and calculates the average of the three annotations. For Turkers, as each annotator may not annotate all the questions, Fleiss's Kappa does not apply, so we only compute the WAWA score. When there is no majority consensus (when all three annotators chose different answers), we randomly picked a consensus answer, and this accounts for 31.45% of the event pairs for Turkers and 19% for trained annotators.

The gold annotations, representing the final decisions, are established through collaborative adjudication meetings with trained annotators. These meetings are facilitated by the paper's first author, who organize and participate in discussions among annotators. The goal of these discussions are to address disagreements and uncertainties, exchange perspectives, and ultimately arrive at a consensus regarding the correct annotations.

## 4.3 Results and Discussion

### 4.3.1 Annotation Agreement and Overall Accuracy

Table 3 provides the accuracy scores for trained annotators, GPT-4, and Turkers on our dataset. The trained annotators achieve the highest accuracy score, closely followed by GPT-4 Zero-Shot, and then Turkers. While GPT-4 underperform trained annotators by a small yet significant margin (64% vs 69.85%), it outperform crowd workers by a large margin (64% vs 42.65%), demonstrating an impressive capability for such a complicated task.

We also measure the inter-annotator agreement (IAA) among human annotators and Turkers in terms of Fleiss' Kappa and WAWA. Comparing the WAWA scores, expert annotators demonstrate significantly higher levels of agreement compared to Turkers. Expert annotators achieved an IAA score of 0.49 in Fleiss' Kappa, falling within a range of scores (0.4 to 0.6) that indicate moderate agreement. Fleiss's Kappa is a chance-corrected metric that is known to be highly stringent. The moderate agreement score suggests a certain level of divergence of opinions or interpretations among the trained annotators, leading to inconsistencies in annotation. This can be attributed to the inherent difficulty and subjectivity involved in CDEC annotation. Events can be described using different tones, intents, levels of granularity, or abstraction, leading to varying interpretations. Furthermore, annotators may possess varying levels of prior knowledge about specific events, resulting in divergent responses when faced with ambiguities. For example, consider the evaluation of coreference between two protests: "esport player Blitzchung's protest leads to his punishment by Blizzard company" and "Blitzchung joined in protest in a video game called Free Hong Kong." Annotators familiar with the Blizzard Hong Kong controversy might immediately tag them as not-related because they are aware of Blitzchung's protest in the Blizzard Hearthstone stream, which is unrelated to the mentioned video game. On the other hand, annotators with no background knowledge may struggle to reason and may either tag it as *cannot-decide* or Identity by making an unsubstantiated inference based on the clues "esport player" and "video game".

### 4.3.2 Accuracy by Topic

Table 4 presents the accuracy scores by topic. Both human annotators and GPT-4 exhibit lowest consistency when annotating event pairs related to the topic "2019 Blizzcon Protest". Upon closer examination, we discover a distinct feature with this topic, characterized by higher trigger word variability and ambiguity. Each cluster under this topic contains a greater variety of unique trigger words. For instance, the event mention cluster representing the action of revoking the reward money taken by Blizzard company from a gamer consists of trigger words such as "revoke", "take", "move", "cancel", "retract", and "act". In addition, we observe that the same trigger word appears in multiple clusters, with a relatively even distribution. For example, the trigger word "protest" appears in clusters that represent the Hong Kong protest on the street, protests organized by gamers in games or online, the specific protest by professional gamer Blitzchung during a Hearthstone live stream, and the protests that occurred in Anaheim Blizzcon in support of Blitzchung. This stands in contrast to easier topics like the Bronx Zoo tiger, where the trigger word "test" primarily refers to the one-time occurrence of the tiger Nadia testing positive for COVID, with a small number of references to other

| Method | Blizzcon | Santos | HCMC | Cyclone | Wildfire | Telescope | Skims | Cruise | Calderwood | Tiger | All Topics |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Trained Annotator | 53.00 | 60.50 | 61.00 | 65.00 | 71.50 | 72.00 | 77.00 | 75.50 | 81.00 | 82.00 | 69.85 |
| GPT-4 | 45.50 | 64.50 | 62.00 | 55.50 | 69.50 | 53.00 | 68.00 | 72.00 | 75.50 | 74.50 | 64.00 |
| Turker | 38.50 | 42.50 | 33.50 | 36.50 | 34.50 | 43.50 | 50.00 | 44.50 | 49.50 | 53.50 | 42.65 |

Table 4: Accuracy Scores by Topic

test events. This highlights the challenges in accurately determining the relation between event mentions when dealing with ambiguous trigger word for both human annotators and GPT-4. In Example (1), based on the context, it can be inferred that the topics and locations of the protests are the same, indicating an “Identity” relation. However, due to the ambiguous nature of the first event mention, humans and GPT-4 made different annotation decisions.

(1) **SENTENCE 1**: Blizzard did not directly address the <protest>, but during the opening ceremony’s keynote speech on Friday, Blizzard president J. Allen Brack said that the company did not handle the situation with blitzchung properly and that he took responsibility for his company’s actions.
**SENTENCE 2**: Messaging in the <protest> at Blizzcon ranged from chants for "Free Hong Kong," to "People over profit" and "Blitzchung did nothing wrong".

Topics Santos and HCMC introduce an additional layer of difficulty related to domain knowledge. Prior to the annotation process, none of our annotators was familiar with the political downfall of the Dos Santos family in Angola or the violations involving certain officials in Ho Chi Minh City. They lack knowledge of the mentioned politicians, and their understanding of political systems is primarily centered around the United States. In contrast, most of our annotators have prior knowledge of figures like Kim Kardashian and possess a common-sense understanding of natural disasters such as cyclones and wildfires, social events like protests against gaming companies or the construction of giant telescopes, and relatively straightforward Covid-related events like Tiger Nadia testing positive or Calderwood’s resignation due to lockdown violations, and the investigation involving Princess Ruby. The accuracy scores show that human annotators performed better in topics they are familiar with. GPT-4 outperformed trained annotators on both these topics that human annotators found challenging.

### 4.3.3 Accuracy by Category

Table 5 presents the distribution of the nine labels in the ground truth annotations, along with the average precision, recall, and F1 score of the three trained annotators. The results indicate that GPT-4 perform comparably to the trained annotators in high frequency labels such as "Identity" and "Not-Related", but show lower performance in low frequency labels. Turkers, on the other hand, demonstrate the lowest performance across all categories. Table 5 shows that the false positive predictions made by GPT-4 for the “Identity” labels are notable, as evidenced by the high recall and low precision of its predictions. GPT-4 tends to exhibit a higher propensity for overreaching and justifying its responses through incorrect inferences, leading to the labeling of relations as “Identity”.

GPT-4 sometimes overlooks the nuanced relations between event pairs in reasoning and tends to label them as “Identity” too readily without looking into all aspects of the event mention, as analyzed in Example (2). Despite capturing the correct action, participants, and location, GPT-4 mislabeled the relation as “Identity” with the true label being “Set-Member”. The discrepancy arises from GPT-4’s failure to recognize the difference in the first event mention, which consists of two instances of visiting (i.e., “the week” and “last”), and in the second event mention, which only refers to one instance of visiting.

(2) **SENTENCE 1**: Scotland’s Chief Medical Officer Catherine Calderwood resigned on Sunday after she broke her own advice to stay at home to help slow the spread of the coronavirus by <visiting> her second home this weekend and last.
**SENTENCE 2**: Dr Calderwood had earlier been given a police warning for breaking the lockdown rules after the Scottish Sun published photographs taken on Saturday of her and her family <visiting> Earlsferry in Fife - more than an hour ’s drive from her main family home in Edinburgh.
**GPT-4 Generation**: The relation between the two marked events in the sentence pair is Identity. Both "visiting" in the first sentence and "visiting" in the second sentence refer to the same action taken by Scotland’s Chief Medical Officer Catherine Calderwood, which is her visiting her second home, thereby breaking lockdown rules.

The most noticeable result from Table 5, however, is that GPT-4 did not label a single instance as “Cannot-Decide”. GPT-4 consistently demonstrates a strong inclination towards making unsupported inferences and, at times, engaging in

| Category | Distribution | Trained Annotators | | | GPT-4 | | | Turkers | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| Identity | 34.00 | 86.90 | 82.94 | 84.88 | 69.17 | 90.74 | 78.50 | 75.05 | 54.41 | 63.09 |
| Concept-Instance | 3.80 | 47.32 | 69.74 | 56.38 | 26.19 | 28.95 | 27.50 | 11.20 | 18.42 | 13.93 |
| Instance-Concept | 2.80 | 32.20 | 67.86 | 43.68 | 5.26 | 3.57 | 4.26 | 5.22 | 10.71 | 7.02 |
| Set-Member | 6.90 | 54.62 | 51.45 | 52.99 | 42.86 | 19.57 | 26.87 | 23.16 | 29.71 | 26.03 |
| Member-Set | 6.95 | 77.89 | 53.24 | 63.25 | 38.00 | 13.67 | 20.11 | 40.27 | 43.17 | 41.67 |
| Whole-Subevent | 3.45 | 24.14 | 30.43 | 26.92 | 19.44 | 30.43 | 23.73 | 2.44 | 5.80 | 3.43 |
| Subevent-whole | 1.85 | 16.28 | 18.92 | 17.50 | 7.62 | 21.62 | 11.27 | 1.75 | 8.11 | 2.89 |
| Not-Related | 35.85 | 76.57 | 76.57 | 76.57 | 85.45 | 78.66 | 81.92 | 70.91 | 48.95 | 57.92 |
| Cannot-Decide | 4.40 | 40.82 | 22.73 | 29.20 | 0.00 | 0.00 | N/A | 3.60 | 4.55 | 4.02 |

Table 5: Label distribution (%) and evaluation results for the different categories.

hallucination. It goes to great lengths to bridge gaps in understanding and actively strives to avoid situations where it cannot provide a definitive answer. In Example (3): based on the available information, it can be deduced that Kim's facial masks have experienced multiple instances of being sold out and restocked. However, there is not enough evidence to conclusively determine if the two event mentions refer to the same restocking occurrence. In this case, GPT-4 incorrectly labeled the relation as "Identity" instead of choosing "Cannot-Decide", failing to acknowledge the ambiguity in the context. Knowing when there is not information to make a decision is a highly complicated process and this could potentially point to fundamental challenge with LLMs. When confronted with a situation where there is insufficient information, LLMs may struggle to accurately recognize and convey their uncertainty. They might attempt to generate a response even if they are not confident or lack the necessary context to provide a reliable answer rather than admitting uncertainty.

(3) **SENTENCE 1**: Kim did not publicly comment on the controversy and has continued to <restock> the collection of facial masks, although it is currently sold out yet again.
**SENTENCE 2**: Claiming that as soon as the brand <restocked> she would donate the generous sum to those affected by the global pandemic , she shared in a statement : " To support Mothers and Children in need during this time , SKIMS is committed to donating $1 M to families affected by COVID-19.
**GPT-4 Generation**: The relation between the two marked events in the sentence pair is Identity. Both sentences refer to the same event of restocking face masks.

Despite the reasoning errors, GPT-4 still display advanced capabilities and broad general knowledge. Occasionally it can discover patterns and insights that may not be immediately apparent to human annotators. In the following Example (4):

(4) **SENTENCE 1**: Kim did not publicly comment on the <controversy> and has continued to restock the collection of facial masks , although it is currently sold out yet again .
**SENTENCE 2**: Criticism on the facial masks is not the first time Kardashian West 's brand has stoked racial <controversy> .
**GPT-4 Generation**: The relation between the two marked events in the sentence pair is Instance-Concept. The first sentence refers to a controversy about a specific event (the selling and restocking of facial masks), while the second sentence refers to a broader concept, encompassing the controversy that Kim Kardashian has generated.

Our human annotators have exhibited confusion when labeling this particular relation as Instance-Concept, often assigning it various other labels. In contrast, GPT-4 accurately identified the second controversy event as a generalization based on the indication of "not the first time", implying that the first controversy event is an instance of the second controversy. However, it is worth noting that although GPT-4 arrived at the correct conclusion, it actually inferred the details of the first controversy incorrectly.

## 5 Conclusion

To address scalability challenges in creating Cross-Document Event Coreference (CDEC) datasets, we explored the feasibility of employing crowdsourcing and GPT-4 using a decontextualized representation of events. Our findings indicate that GPT-4 outperforms crowd workers by a large margin and shows comparable performance to trained annotators. We also observe variations in performance across different topics and individual coreference categories and uncovered issues related to reasoning and hallucination in GPT-4's performance in the CDEC annotation task. Despite its limitations, our work suggests that GPT-4 has the potential to replace human effort in creating complex and labor intensive CDEC datasets in at least some settings at scale. Given the far superior performance of GPT-4 over crowd workers, it no longer makes sense to resort to untrained annotators in crowdsourcing settings for such complex annotation tasks. The best approach might be one that combines the strengths of LLMs such as GPT-4 with highly trained annotators. We leave it to future work as to exactly how that combination should work.

## Acknowledgement

This research was supported by the CNS Division of the National Science Foundation (Awards No. 2213804, 2213805). We express our gratitude for their financial backing, which was pivotal in advancing our understanding of cross-document event coreference.

We also wish to extend our appreciation to Cloudbank, which supplied indispensable computational resources throughout our project.

Additionally, our thanks go to the diligent team of annotators. Their invaluable time and expertise were instrumental in curating the high-quality dataset pivotal to this study, and we recognize their significant contribution to our project.

## References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Michael Bugert and Iryna Gurevych. 2021. Event coreference data (almost) for free: Mining hyperlinks from online news. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 471–491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

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).

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. 2016. Building a cross-document event-event relation corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*. ACM.

Chao-Yi Lu and Sin-En Lu. 2021. A survey of approaches to automatic question generation:from 2019 to early 2021. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 151–162, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2019. Controlled crowdsourcing for high-quality qa-srl annotation.

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).

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiarui Yao, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. Factuality assessment as modal dependency parsing. 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 1540–1550, Online. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks.

# A Appendix

## A.1 Limitations

This paper utilized decontextualization and event extraction to select and formulate potential event pairs. Evaluating the errors arising from these methods, and their potential impact on both GPT-4 and human annotations, is reserved for future research.

Decontextualization goes beyond mere simplification as it involves restating a sentence together with its contextual information in a way that allows it to be understood independently of its original context while preserving its intended meaning. The goal of decontextualization is to capture the relevant contextual details and integrate them into a single sentence without sacrificing any crucial information. However, in practice, our current decontextualization model falls short of consistently providing useful results. At times, it may engage in unnecessary noun phrase swappings or insertions that adds little value. Furthermore, there is a risk of errors occurring during the decontextualization process, which can potentially propagate and impact downstream tasks. In the following example, the decontextualization model hallucinates a false context for the death of the saplings. Annotators with prior knowledge about the HS2 project can readily identify the absurdity of the context and recognize it as a decontextualization error. However, annotators without the relevant background knowledge may mistakenly assume the described battle to be true. Consequently this wrong assumption will propagate to downstream event coreference task.

> **Original**: Up to 350,000 saplings have so far been planted near the £ 56bn train line , but two Warwickshire farmers think up to 80% on their land have died .
> **Decontexualized**: Up to 350,000 saplings have so far been planted near the £ 56bn train line , but two Warwickshire farmers think up to 80% on their land have died *in the Battle of High Speed Rail 2 ( HS2 )* .

Our open-domain event detection model sometimes identified false positive event triggers, leading to potential ambiguity for our annotators and influencing their judgments. For the following example, our model tagged <accused> as an event trigger. Yet, this term refers to the individuals facing accusations rather than the act of accusation. While <accused> here isn't an event in itself, it implies a related event: the court's act of accusing dos Santos, Dokolo, and da Silva. Consequently, some annotators missed this extraction error and incorrectly linked it to other accusation events.

> The asset freeze applies to personal bank accounts of dos Santos, Dokolo and da Silva in Angola and stakes they hold in Angolan firms including Unitel, BFA and ZAP MIDIA, and the order said the central bank would ensure that no funds leave the personal bank accounts of the three <accused>.

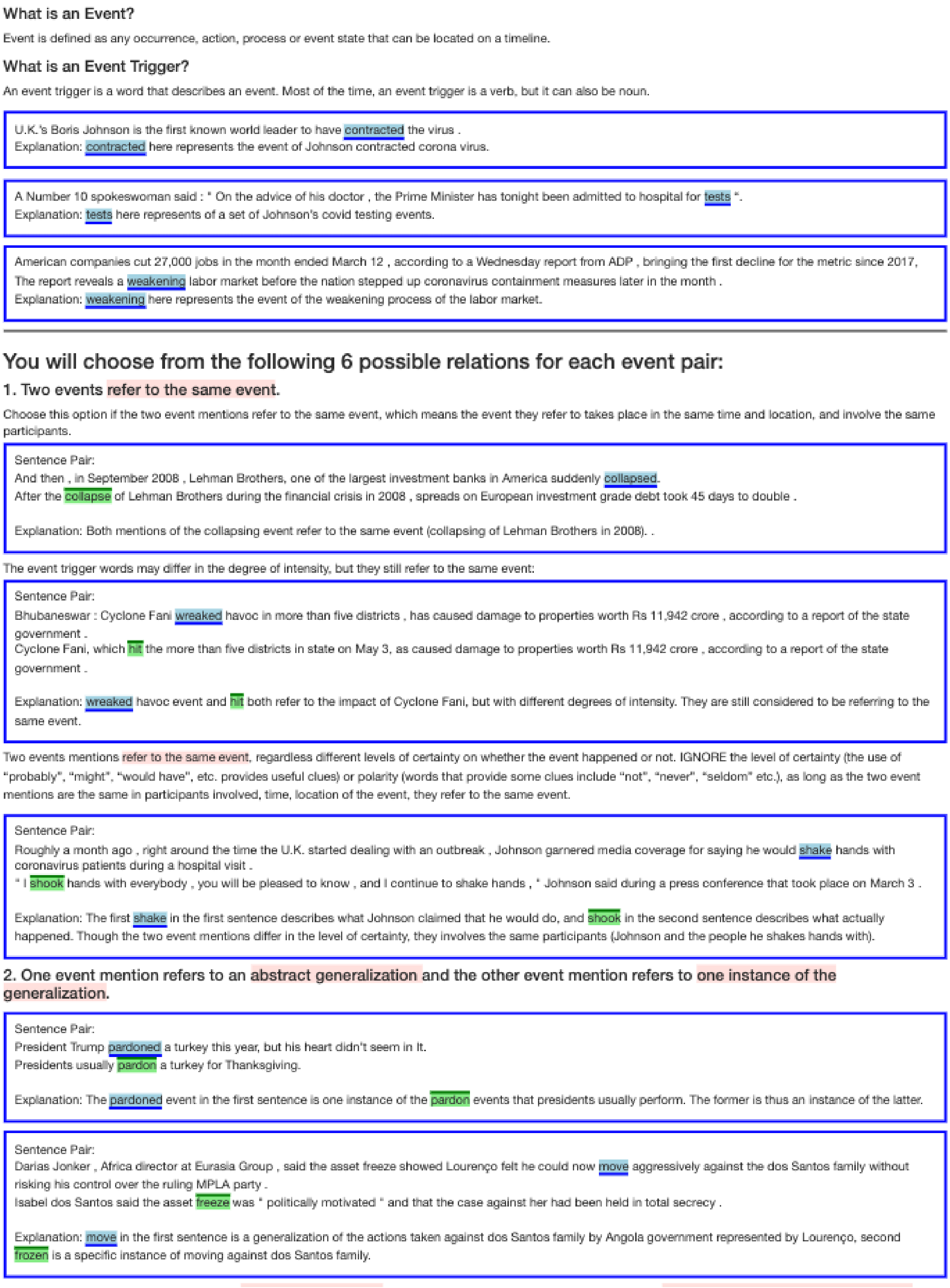
We selected GPT-4 due to its state-of-the-art performance and its adeptness at handling reasoning and language comprehension tasks. Future studies should evaluate how its training data, as well as any inherent biases or specialties, might influence cross-document event coreference results, and further validate our findings using different language models.

## A.2 Annotation Guidelines

See Figure 2 and Figure 3.

## A.3 Data Creation Pipeline

See Figure 4.



### What is an Event?

Event is defined as any occurrence, action, process or event state that can be located on a timeline.

### What is an Event Trigger?

An event trigger is a word that describes an event. Most of the time, an event trigger is a verb, but it can also be noun.

U.K.'s Boris Johnson is the first known world leader to have contracted the virus .
Explanation: contracted here represents the event of Johnson contracted corona virus.

A Number 10 spokeswoman said : " On the advice of his doctor , the Prime Minister has tonight been admitted to hospital for tests ".
Explanation: tests here represents of a set of Johnson's covid testing events.

American companies cut 27,000 jobs in the month ended March 12 , according to a Wednesday report from ADP , bringing the first decline for the metric since 2017, The report reveals a weakening labor market before the nation stepped up coronavirus containment measures later in the month .
Explanation: weakening here represents the event of the weakening process of the labor market.

---

## You will choose from the following 6 possible relations for each event pair:

### 1. Two events refer to the same event.

Choose this option if the two event mentions refer to the same event, which means the event they refer to takes place in the same time and location, and involve the same participants.

Sentence Pair:
And then , in September 2008 , Lehman Brothers, one of the largest investment banks in America suddenly collapsed.
After the collapse of Lehman Brothers during the financial crisis in 2008 , spreads on European investment grade debt took 45 days to double .

Explanation: Both mentions of the collapsing event refer to the same event (collapsing of Lehman Brothers in 2008). .

The event trigger words may differ in the degree of intensity, but they still refer to the same event:

Sentence Pair:
Bhubaneswar : Cyclone Fani wreaked havoc in more than five districts , has caused damage to properties worth Rs 11,942 crore , according to a report of the state government .
Cyclone Fani, which hit the more than five districts in state on May 3, as caused damage to properties worth Rs 11,942 crore , according to a report of the state government .

Explanation: wreaked havoc event and hit both refer to the impact of Cyclone Fani, but with different degrees of intensity. They are still considered to be referring to the same event.

Two events mentions refer to the same event, regardless different levels of certainty on whether the event happened or not. IGNORE the level of certainty (the use of "probably", "might", "would have", etc. provides useful clues) or polarity (words that provide some clues include "not", "never", "seldom" etc.), as long as the two event mentions are the same in participants involved, time, location of the event, they refer to the same event.

Sentence Pair:
Roughly a month ago , right around the time the U.K. started dealing with an outbreak , Johnson garnered media coverage for saying he would shake hands with coronavirus patients during a hospital visit .
" I shook hands with everybody , you will be pleased to know , and I continue to shake hands , " Johnson said during a press conference that took place on March 3 .

Explanation: The first shake in the first sentence describes what Johnson claimed that he would do, and shook in the second sentence describes what actually happened. Though the two event mentions differ in the level of certainty, they involves the same participants (Johnson and the people he shakes hands with).

### 2. One event mention refers to an abstract generalization and the other event mention refers to one instance of the generalization.

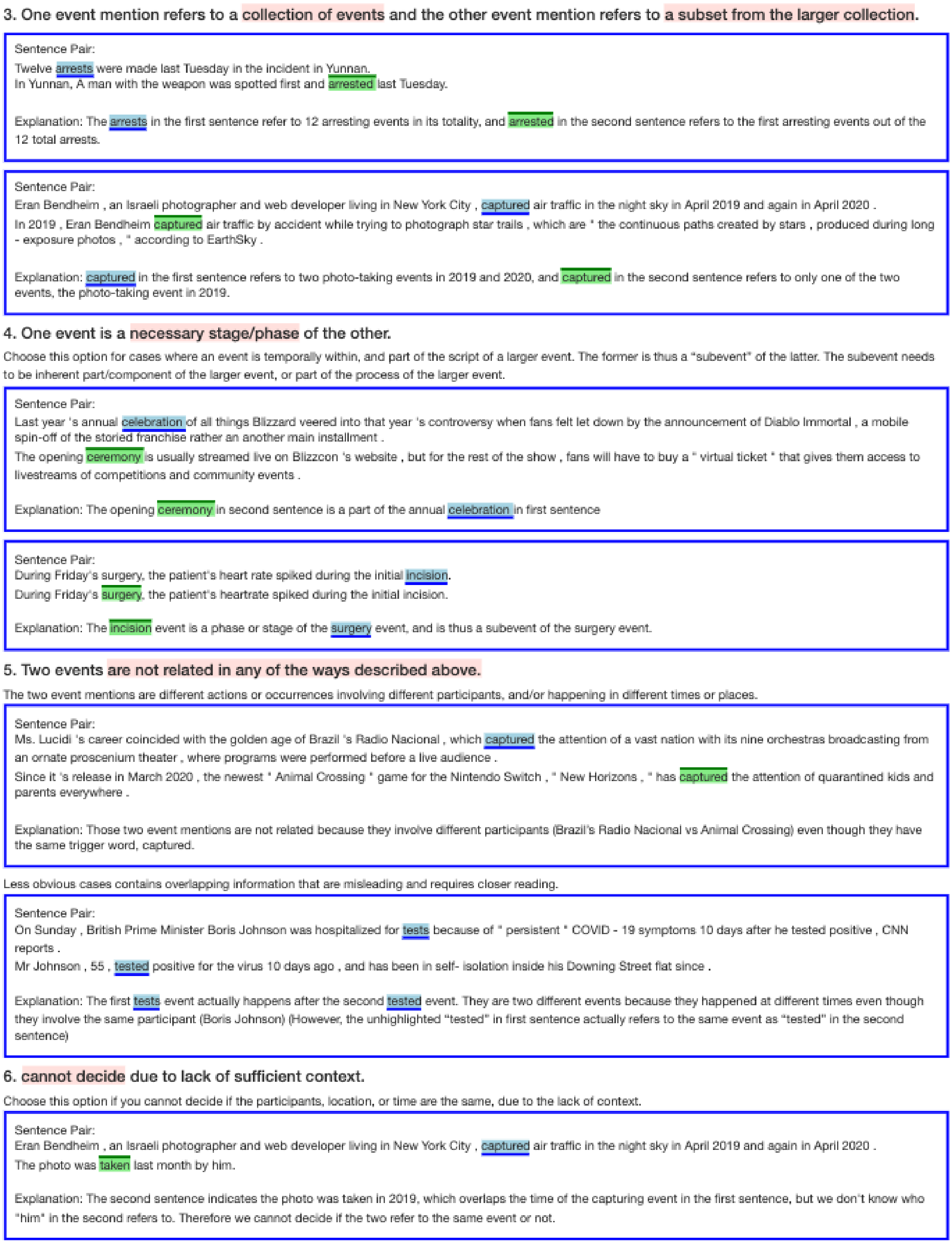
Sentence Pair:
President Trump pardoned a turkey this year, but his heart didn't seem in it.
Presidents usually pardon a turkey for Thanksgiving.

Explanation: The pardoned event in the first sentence is one instance of the pardon events that presidents usually perform. The former is thus an instance of the latter.

Sentence Pair:
Darias Jonker , Africa director at Eurasia Group , said the asset freeze showed Lourenço felt he could now move aggressively against the dos Santos family without risking his control over the ruling MPLA party .
Isabel dos Santos said the asset freeze was " politically motivated " and that the case against her had been held in total secrecy .

Explanation: move in the first sentence is a generalization of the actions taken against dos Santos family by Angola government represented by Lourenço, second frozen is a specific instance of moving against dos Santos family.

Figure 2: Annotation Guidelines(Part1/2).



3. One event mention refers to a collection of events and the other event mention refers to a subset from the larger collection.

Sentence Pair:
Twelve arrests were made last Tuesday in the incident in Yunnan.
In Yunnan, A man with the weapon was spotted first and arrested last Tuesday.

Explanation: The arrests in the first sentence refer to 12 arresting events in its totality, and arrested in the second sentence refers to the first arresting events out of the 12 total arrests.

Sentence Pair:
Eran Bendheim , an Israeli photographer and web developer living in New York City , captured air traffic in the night sky in April 2019 and again in April 2020 .
In 2019 , Eran Bendheim captured air traffic by accident while trying to photograph star trails , which are " the continuous paths created by stars , produced during long - exposure photos , " according to EarthSky .

Explanation: captured in the first sentence refers to two photo-taking events in 2019 and 2020, and captured in the second sentence refers to only one of the two events, the photo-taking event in 2019.

4. One event is a necessary stage/phase of the other.

Choose this option for cases where an event is temporally within, and part of the script of a larger event. The former is thus a "subevent" of the latter. The subevent needs to be inherent part/component of the larger event, or part of the process of the larger event.

Sentence Pair:
Last year 's annual celebration of all things Blizzard veered into that year 's controversy when fans felt let down by the announcement of Diablo Immortal , a mobile spin-off of the storied franchise rather an another main installment .
The opening ceremony is usually streamed live on Blizzcon 's website , but for the rest of the show , fans will have to buy a " virtual ticket " that gives them access to livestreams of competitions and community events .

Explanation: The opening ceremony in second sentence is a part of the annual celebration in first sentence

Sentence Pair:
During Friday's surgery, the patient's heart rate spiked during the initial incision.
During Friday's surgery, the patient's heartrate spiked during the initial incision.

Explanation: The incision event is a phase or stage of the surgery event, and is thus a subevent of the surgery event.

5. Two events are not related in any of the ways described above.

The two event mentions are different actions or occurrences involving different participants, and/or happening in different times or places.

Sentence Pair:
Ms. Lucidi 's career coincided with the golden age of Brazil 's Radio Nacional , which captured the attention of a vast nation with its nine orchestras broadcasting from an ornate proscenium theater , where programs were performed before a live audience .
Since it 's release in March 2020 , the newest " Animal Crossing " game for the Nintendo Switch , " New Horizons , " has captured the attention of quarantined kids and parents everywhere .

Explanation: Those two event mentions are not related because they involve different participants (Brazil's Radio Nacional vs Animal Crossing) even though they have the same trigger word, captured.

Less obvious cases contains overlapping information that are misleading and requires closer reading.

Sentence Pair:
On Sunday , British Prime Minister Boris Johnson was hospitalized for tests because of " persistent " COVID - 19 symptoms 10 days after he tested positive , CNN reports .
Mr Johnson , 55 , tested positive for the virus 10 days ago , and has been in self- isolation inside his Downing Street flat since .

Explanation: The first tests event actually happens after the second tested event. They are two different events because they happened at different times even though they involve the same participant (Boris Johnson) (However, the unhighlighted "tested" in first sentence actually refers to the same event as "tested" in the second sentence)

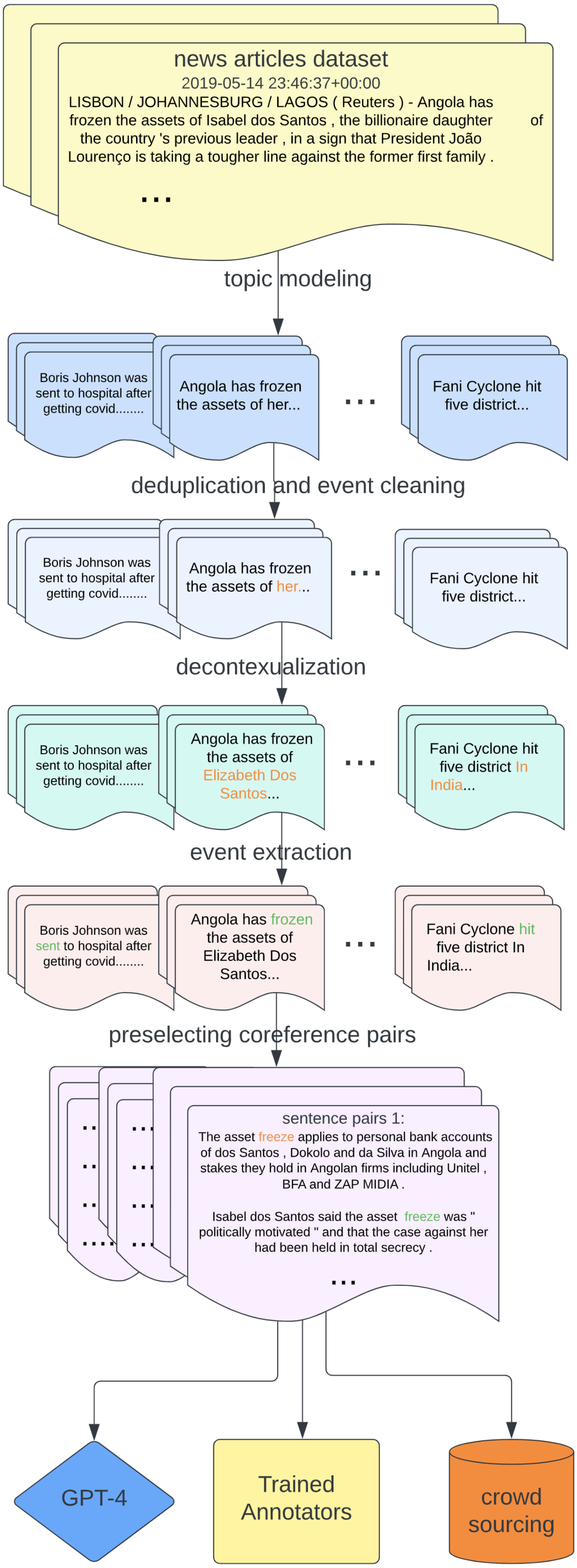
6. cannot decide due to lack of sufficient context.

Choose this option if you cannot decide if the participants, location, or time are the same, due to the lack of context.

Sentence Pair:
Eran Bendheim , an Israeli photographer and web developer living in New York City , captured air traffic in the night sky in April 2019 and again in April 2020 .
The photo was taken last month by him.

Explanation: The second sentence indicates the photo was taken in 2019, which overlaps the time of the capturing event in the first sentence, but we don't know who "him" in the second refers to. Therefore we cannot decide if the two refer to the same event or not.

Figure 3: Annotation Guidelines(Part2/2).



news articles dataset

2019-05-14 23:46:37+00:00

LISBON / JOHANNESBURG / LAGOS ( Reuters ) - Angola has frozen the assets of Isabel dos Santos , the billionaire daughter of the country 's previous leader , in a sign that President João Lourenço is taking a tougher line against the former first family .

...

sentence pairs 1:

The asset freeze applies to personal bank accounts of dos Santos , Dokolo and da Silva in Angola and stakes they hold in Angolan firms including Unitel , BFA and ZAP MIDIA .

Isabel dos Santos said the asset freeze was " politically motivated " and that the case against her had been held in total secrecy .

...

topic modeling
Boris Johnson was sent to hospital after getting covid........
Angola has frozen the assets of her...
...
Fani Cyclone hit five district...
deduplication and event cleaning
Boris Johnson was sent to hospital after getting covid........
Angola has frozen the assets of her...
...
Fani Cyclone hit five district...
decontexualization
Boris Johnson was sent to hospital after getting covid........
Angola has frozen the assets of Elizabeth Dos Santos...
...
Fani Cyclone hit five district In India...
event extraction
Boris Johnson was sent to hospital after getting covid........
Angola has frozen the assets of Elizabeth Dos Santos...
...
Fani Cyclone hit five district In India...
preselecting coreference pairs
GPT-4
Trained Annotators
crowd sourcing

Figure 4: Data Creation Pipeline.