LegalCore: A Dataset for Event *Core*ference Resolution in *Legal* Documents

Kangda Wei[1], Xi Shi[1], Jonathan Tong[1], Sai Ramana Reddy[1], Anandhavelu Natarajan[2], Rajiv Jain[2], Aparna Garimella[2], Ruihong Huang[1]
[1]Department of Computer Science and Engineering, Texas A&M University, College Station, TX
[2]Adobe Research
{kangda, xishi, tongjo, sairamana, huangrh}@tamu.edu
{anandvn, rajijain, garimell}@adobe.com

Abstract

Recognizing events and their coreferential mentions in a document is essential for understanding semantic meanings of text. The existing research on event coreference resolution is mostly limited to news articles. In this paper, we present the first dataset for the legal domain, LegalCore, which has been annotated with comprehensive event and event coreference information. The legal contract documents we annotated in this dataset are several times longer than news articles, with an average length of around 25k tokens per document. The annotations show that legal documents have dense event mentions and feature both short-distance and super long-distance coreference links between event mentions. We further benchmark mainstream Large Language Models (LLMs) on this dataset for both event identification and event coreference resolution tasks, and find that this dataset poses significant challenges for both open-source and proprietary LLMs, which all perform significantly worse than a supervised baseline. Our data and code are available at https://github.com/WeiKangda/LegalCore

1 Introduction

Identifying event mentions and grouping event mentions based on their coreference relations is a key step for text semantic understanding and necessary for further event structure analysis. However, research on event coreference resolution is mostly limited to news articles as the annotated datasets for event coreference resolution are mostly for this domain (Walker and Consortium, 2005; Cybulska and Vossen, 2014; Ellis et al., 2015, 2016; Getman et al., 2017; Wang et al., 2022). There are a few datasets for other more specific domains, such as Twitter (Ritter et al., 2012), literature(Sims et al., 2019) and biomedical texts(Thompson et al., 2009), but these datasets often annotate individual event mentions only and lack event coreference information.

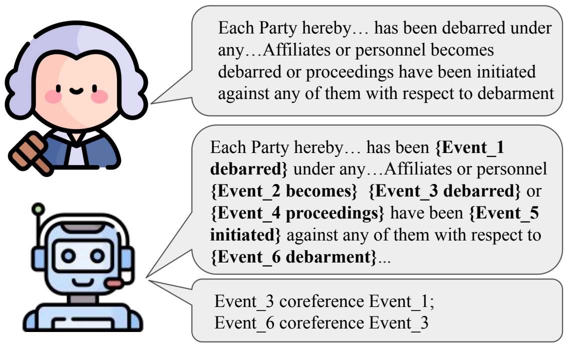

Figure 1: Event detection and event coreference resolution for a legal text excerpt.

In this paper, we present the first dataset for the legal domain, LegalCore, which has been annotated with comprehensive event and event coreference information. The legal contract documents in this dataset are several times longer than news articles, with an average length of around 2.5k tokens per document. LegalCore contains 100 legal contract documents and around 250k tokens, comparable in size to the main event-annotated datasets, such as ACE 2005 (Walker and Consortium, 2005) and ECB+ (Cybulska and Vossen, 2014).

As illustrated in Figure 2, we first annotate individual event words as event mentions in each document, without constraints on event type. As shown by the document excerpt example, legal documents have dense event mentions, and there are 23,183 event mentions annotated in total in our dataset, roughly one event word in every ten tokens of a legal document.

Next, we annotate event coreference relations. But as each legal document is so long and usually contains a brief introductory section followed by a series of numbered sections, it is hard to identify all the coreferential mentions of an event by going through the document once. As shown in the example document of Figure 2, the beginning section in-

Findings of the Association for Computational Linguistics: ACL 2025, pages 25044–25059
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

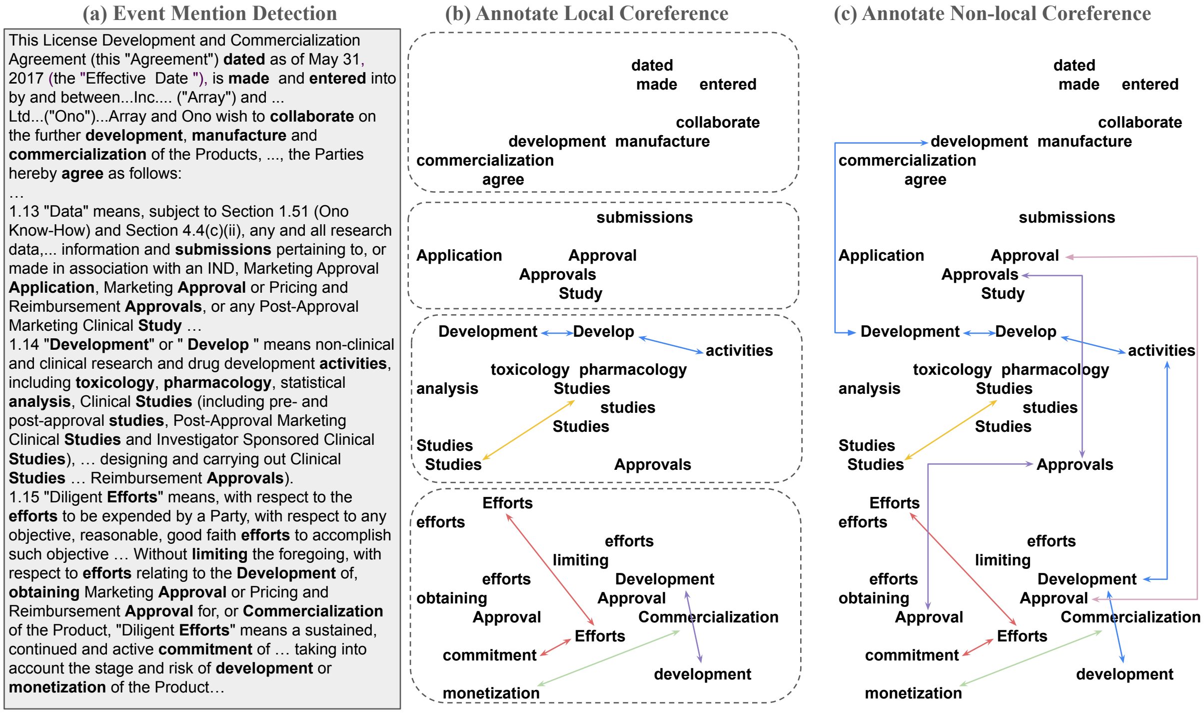

Figure 2: Illustration of the 3-stages annotation process for `LegalCore`. First, event mentions are annotated in legal documents. Next, local coreference links are identified within each section. Finally, non-local (cross-section) coreference links are annotated. **Bold**: event mentions. **Arrows**: coreference links.

troduces the main involved parties and the theme of the contract, and the following numbered sections further specify rules or elaborate on the main terminology used in the contract. Each section often cites other sections for references. Therefore, we design a two-pass procedure for annotating event coreference relations and recover complete event clusters in a hierarchical fashion, where we first annotate local event coreference relations within each section and then annotate non-local event coreference relations spanning across sections.

After completing both passes of event coreference annotations, there are 853 event clusters identified in our datasets, where 653 event clusters are local clusters and only contain event mentions within one section, and the remaining 200 event clusters contain event mentions from two or more sections. Among the 200 non-local event clusters, a little over half of them span across two sections and the remaining span across three or more sections, in particular, about 30 (15%) non-local event clusters span across six or more sections. Accordingly, the legal documents feature both short-distance and super long-distance coreference links between event mentions. If we measure the distance of coreference links between an event mention and its nearest antecedent mention as the number of tokens between the two event mentions, we observe over half of such coreference links have a distance less than 50 tokens, meanwhile, we observe a significant portion of super long coreference links covering more than 600 or even more than 1000 tokens.

In addition, with the increasing popularity of LLM, we benchmark the performance of the main LLMs in `LegalCore`, including open-source and proprietary ones, such as Llama-3.1 (Grattafiori et al., 2024) and GPT-4 (OpenAI et al., 2024). Despite strong performance of LLMs in many tasks (Chang et al., 2023; Du et al., 2024; Wei et al., 2023a,b, 2022), our experiments show that `LegalCore` poses significant challenges for both open-source and proprietary LLMs, and the performance of LLMs on both event identification and event coreference resolution are still worse than a supervised baseline.

To summarize, our contributions are mainly two:

- We introduce `LegalCore`, the first dataset for the legal domain that has been annotated with comprehensive event and event coreference information.
- We benchmark mainstream LLMs performance on `LegalCore`, finding that both event detection and event coreference remain challenging to LLMs.

2 Related Works

Event Identification There are many existing datasets annotated with event information, some require the models to identify event mentions and classify them into specific event types (Ellis et al., 2015, 2016; Getman et al., 2017; Wang et al., 2020; Walker and Consortium, 2005), while others require the models to extract event mentions of all types (Allan, 2002; Minard et al., 2016; Araki and Mitamura, 2018; Sims et al., 2019; Liu et al., 2019a), and `LegalCore` falls into the later category. Most of the previous works (Walker and Consortium, 2005; Ellis et al., 2015, 2016; Getman et al., 2017; Cybulska and Vossen, 2014; Pustejovsky et al., 2003) focus on news domains, but a few datasets have been developed for other specific domains, for example Twitter (Ritter et al., 2012; Guo et al., 2013; Chen et al., 2022), literature (Sims et al., 2019), finance (Chen et al., 2021), and biomedical texts (Pyysalo et al., 2007; Kim et al., 2007; Thompson et al., 2009; Buyko et al., 2010; Nédellec et al., 2013), but none of the previous datasets consider the legal domain. To the best of our knowledge, `LegalCore` is the first dataset for the legal domain annotated with event information.

Event Coreference Resolution Event Coreference Resolution is a key and challenging NLP task, and many event coreference datasets have been constructed. Previous datasets for event coreference are mostly based on news articles (Ellis et al., 2015, 2016; Getman et al., 2017; Cybulska and Vossen, 2014; Pradhan et al., 2007). The most recent large dataset, MAVEN-ERE (Wang et al., 2022), contains annotations of event coreference relations as well as other event relations such as temporal relations and causal relations, but the annotated Wikipedia articles are still mostly news documents. The event coreference relations annotated in `LegalCore` are expected to enable more studies on event coreference resolution for the legal domain and be highly valuable for developing real-world applications for this important domain.

3 Dataset Construction

The `LegalCore` dataset contains 100 legal contract documents that were selected from the CUAD dataset (Hendrycks et al., 2021), a public dataset used for identifying key clauses in legal contracts. We will publish our annotated dataset.

The dataset construction is done in three phases, namely annotating event mentions, identifying local coreference links, and identifying non-local coreference links, as shown in Figure 2. We follow the annotation guidelines of O'Gorman et al. (2016) for performing the first two phases of annotations, we create our own annotation guidelines for annotating non-local coreference links, and we will publish our data annotation guidelines.

3.1 Event Mentions Annotation

Task Description We first annotate event mentions in each legal contract document. We define an event as any occurrence, action, process, or state that belongs on a timeline and can take various syntactic forms, including verbs, nominalizations, nouns, or adjectives, as outlined by O'Gorman et al. (2016).

Annotation We follow the annotation guidelines of O'Gorman et al. (2016) and have two annotators annotate event mentions. Given the plain text of a legal document, the annotators are asked to identify all event mentions that occur in the document. The annotators are trained and instructed that event determination should rely solely on semantics—whether something belongs on a timeline. Syntactic form is secondary and considered later as an event can take any syntactic realization. At this stage, we only focus on the semantic aspect and determine whether the words represent changes, transitions, or states occurring in the world. To ensure the quality of the annotated data, we sample five documents and asked both annotators to identify all the event mentions. The inter-annotator's agreement is 80.2% (Cohen's kappa). In total, we have 23,183 event mentions annotated in this dataset.

3.2 Local Coreference Annotation

Task Description Event coreference relations link event mentions referring to the same event in space and time. The coreference relation is both symmetrical and transitive. In this stage, the annotators are only asked to annotated local coreference relations. A coreference link is considered local if both event mentions are within the same section of a legal document. As shown in the example of Figure 2, sections in a legal document usually have a section number, like *3. Payment*.

Annotation We follow the annotation guidelines of O'Gorman et al. (2016) and have two trained annotators annotate local event coreference. Given a legal document with annotated event mentions, the

	# Events	# Mentions per event
Local	653	2.5
Non-local	200	4.4

Table 1: Statistics of Non-singleton Events.

annotators are asked to identify coreference links within the same section. An event mention should only be linked to its nearest antecedent mention if any. To ensure the quality of the annotated data, we have the two annotators annotate five common documents and measure the inter-annotator agreement. The Cohen's kappa is 70.0% for this stage of annotations.

3.3 Non-local Coreference Annotation

Task Description For this stage, the annotators are only asked to annotated non-local coreference relations. A coreference link is considered non-local if the two event mentions are within different sections of a document. Non-local coreference links are essentially cross-section links.

Annotation We create our own annotation guidelines and have two annotators annotate non-local event coreference relations. Given a legal document with annotated event mentions, the annotators are asked to identify coreference links cross different sections. We do not explicitly instruct annotators to exclude mentions that are already part of local clusters when identifying non-local coreference links. Instead, the local and non-local clusters naturally emerge after completing the two-stage coreference annotation process. To ensure the quality of the annotated data, we sample five documents and asked both annotators identify cross-section coreference links. The inter-annotator agreement is 74.8% (Cohen's kappa).

3.4 Basic Statistics of Non-singleton Events

As shown in table 1, we have 853 non-singleton events in the dataset. Among them, 653 event clusters are local coreference clusters with an average of 2.5 event mentions per cluster. For the remaining 200 non-local coreference clusters with an average of 4.4 mentions per cluster. Overall, the non-singleton events have 2.9 mentions per event.

Figure 3 further shows the distribution of non-local coreference clusters based on the number of sections each cluster span across. A little over half of the non-local coreference clusters span across two sections and the remaining span across three or more sections, in particular, about 30 (15%)

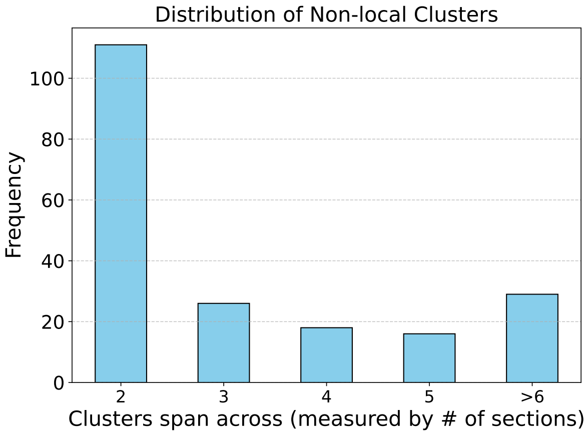

Figure 3: Histogram of Non-Local Cluster Distribution: This chart illustrates the distribution of non-local clusters based on the number of sections they span.

Domain	Dataset	#Doc	Total #Tokens	#Tokens per Doc	#Mention per Doc
News	ACE 2005	599	294,857	506	9
	ECB+	982	362,546	369	15
	TAC KBP	1,075	694,540	646	27
	MAVEN-ERE	4480	1,275,644	285	25
News & Forum Discussion	RED	95	54,287	571	92
Legal	`LegalCore`	100	249,523	2495	232

Table 2: Statistics of event coreference relations in `LegalCore` and existing datasets. Notice that `LegalCore` has the largest # of Tokens per Doc, four times than the dataset in the second place, and the largest # of Mentions per Doc, over twice than the dataset in the second place.

non-local event clusters span across six or more sections, and these long span coreference clusters can pose greater challenges to event coreference resolution.

4 Dataset Analysis

4.1 Statistics Comparing to Existing Datasets

Table 2 compares the size of `LegalCore` with existing widely-used event coreference datasets, including ACE 2005 (Walker and Consortium, 2005), ECB+ (Cybulska and Vossen, 2014), TAC KBP, MAVEN-ERE (Wang et al., 2022), and RED (O'Gorman et al., 2016). We can see that most datasets contain event coreference annotations for news articles only, and only RED contains a small number of other types of texts. Considering the total number of tokens, `LegalCore` is comparable in size to the commonly used dataset ACE 2005 (Walker and Consortium, 2005) and ECB+ (Cybulska and Vossen, 2014). TAC KBP shown here consists of a collection of smaller datasets, TAC

Dataset	< 50 (%)	50 - 200(%)	> 200 (%)	Average
ACE 2005	36.4	27.9	35.6	192
TAC KBP	22.8	26.5	50.7	536
MAVEN-ERE	31.9	49.4	18.8	122
LegalCore	55.7	25.2	19.1	158

Table 3: The distributions and average values of distances (measured in #tokens) of coreference links for LegalCore and existing datasets.

KBP 2015 (Ellis et al., 2015), 2016(Ellis et al., 2016), 2017(Getman et al., 2017) and two other LDC datasets[1], following previous works (Wang et al., 2022; Lu and Ng, 2021a,b). MAVEN-ERE (Wang et al., 2022) is broad coverage news dataset covering 168 event types and is significantly bigger than all the other datasets.

Compared to all the existing datasets, the legal documents in LegalCore are several times longer and contain significant more event mentions per document.

4.2 Distance between Coreferential Mentions

Recognizing relations between distant event mention pairs is crucial for discourse-level document comprehension (Naik et al., 2019), yet capturing long-range dependencies remains a persistent challenge for NLP models. Therefore, we examine the distance distributions of annotated event relations in LegalCore and compare them with the most widely used existing datasets in Table 3. The distance of a coreference link is measured as the number of tokens between an event mention and its nearest antecedent mention.

Interestingly, despite having much longer document, LegalCore has similar average distances and much more coreference event links with short distance (< 50 tokens) comparing to previous datasets. This is mainly because of two reasons: First, legal documents often reference the same event multiple times within a section for clarity and consistency. Lawyers typically avoid using synonyms, instead opting for exact wording to ensure precision and reduce ambiguity. As a result, event mentions referring to the same event tend to appear close together in order to reinforce the document's accuracy and legal integrity. Second, LegalCore is annotated with denser event mentions, averaging 10.8 tokens between mentions compared to 24.9 tokens in TAC KBP. This dense event mention annotation, combined with our thorough event coreference annotation schema—which annotates local event coreference before addressing non-local cases —results in most event coreference links occurring over short distances.

[1] LDC2015E29 and LDC2015E68.

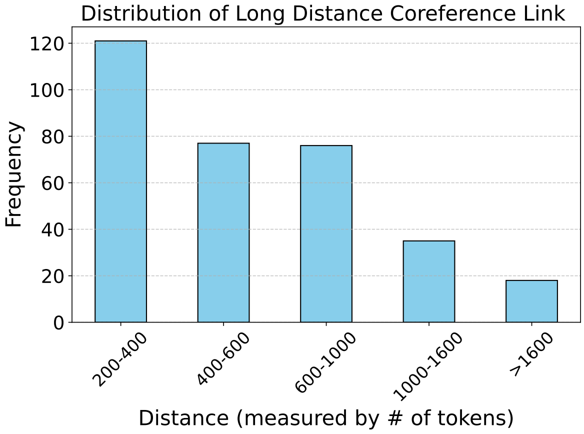

Figure 4: Histogram of long distance coreference links. A coreference link is considered long if the distance measured by # of tokens between two event mentions is over 200 tokens.

On the other hand, our dataset contains a significant number of super long-distance coreference links. We zoom in the long-distance coreference links where the event mentions are over 200 tokens apart, and show the histogram of long distance coreference links in Figure 4. We can see that a significant number of super long-distance coreference links, where the two event mentions have over 400 tokens or even over 600 tokens in between, dozens of coreference links have even over 1000 tokens in between. These coreference links with a super long distance pose great challenges to event coreference resolution models.

5 Experiments and Analysis

5.1 Experiment Setup

Benchmark LLMs We also benchmark the performance of mainstream LLMs on LegalCore. The following models are used in the different event-related tasks: Llama-3.1 (Grattafiori et al., 2024), Mistral-Nemo (Mistral AI, 2024), Qwen-2.5 (Qwen et al., 2025), and GPT-4 (OpenAI et al., 2024). The exact model versions can be found in Appendix A.2.

For event identification, we ask LLMs to annotate event mentions sentence by sentence. For event coreference, we ask LLMs to find all event mentions at once due to the existence of non-local coreference chains, as one mention can appear at the beginning of a document while the other mention appears at the very end. The prompts used for

LLM benchmarking can be found in Appendix A.3.

We also evaluate LLMs in an end-to-end fashion and report the final performance, where we perform event coreference resolution on the model identified noisy event mentions.

Input	**Output**
JSC NOC KazakhOil, hereinafter referred to as the "Company", in the person of Executive Marketing Director Ms. A. M. Rakhimbekov, acting on the basis of the Power of Attorney (1) 1-13 dated January 3, 2000.	JSC NOC KazakhOil, hereinafter {E1 referred} to as the "Company", in the person of Executive Marketing Director Ms. A. M. Rakhimbekov, {E2 acting} on the basis of the Power of Attorney (1) 1-13 {E3 dated} January 3, 2000.

Table 4: Example of an input and output pair for fine-tuning the T-5 model used for supervised event detection.

Supervised Baseline We built a supervised baseline for both event identification and event coreference resolution. For event identification, we refer to Hicke and Mimno (2024) and fine-tune T-5 models (Raffel et al., 2023) to take a raw sentence as the input and output the same sentence marked with event mentions. Table 4 shows an example of the input and output pair. We experimented with T-5 models of different sizes: small, base, and large.

For event coreference resolution, we follow Wang et al. (2022) and adopt a pre-trained RoBERTa-base model (Liu et al., 2019b) for identifying event coreference relations. Firstly, the whole document is encoded using RoBERTa-base. For documents longer than RoBERTa-base's context window, we split them into chunks and encode each chunk separately, following the approach outlined by Wang et al. (2022). Next, the contextualized representations at the positions of each event mentions are extracted from the encoded document. Finally. the extracted representations are then fed into a classification head in a pair-wise fashion to determine wether if there is a coreference link exists between the two event mentions. We also report the end-to-end performance for the supervised baseline. All the experiments of the supervised baseline were conducted using 5-fold cross-validation.

Metrics For event identification, we adopt the standard micro-averaged precision, recall, and F-1 metrics. Following previous works (Wei et al., 2024; Wang et al., 2022; Choubey and Huang, 2017), we evaluate event coreference resolution performance by adopting MUC(Vilain et al., 1995), B^3(Bagga and Baldwin, 1998), CEAF$_e$(Luo, 2005), and BLANC (Recasens and Hovy, 2011) metrics.

	Model	Precision	Recall	F-1
LLMs	GPT-4	72.5	39.0	50.7
	Llama-3.1	52.3	47.7	49.9
	Mistral-Nemo	66.5	37.4	47.9
	Qwen-2.5	83.4	48.7	61.5
Supervised	t5-small	98.7	98.0	98.4
	t5-base	99.1	98.6	**98.9**
	t5-large	99.0	98.5	98.8

Table 5: Event Detection Performance. Supervised baseline almost reach perfect score. LLMs are tested with zero-shot setting and significantly underperform the supervised-baseline.

5.2 Event Identification Results

	Setting	Precision	Recall	F-1
GPT-4	Zero-shot	72.5	39.0	50.7
	One-shot	70.9	79.6	75.0
	Two-shot	79.6	75.9	77.7
Llama-3.1	Zero-shot	52.3	47.7	49.9
	One-shot	47.3	70.8	56.7
	Two-shot	51.5	68.7	58.9
Mistral-Nemo	Zero-shot	66.5	37.4	47.9
	One-shot	73.1	40.8	52.4
	Two-shot	67.2	45.5	54.3
Qwen-2.5	Zero-shot	83.4	48.7	61.5
	One-shot	80.4	54.4	64.9
	Two-shot	82.8	47.6	60.5

Table 6: Event Detection Performance for LLMs with Few-shot Prompting. Recall are improved significantly comparing to zero-shot setting.

Table 5 shows that all LLMs significantly underperform the supervised baseline, and the latter performs almost perfectly on event identification. Furthermore, the results suggest that the primary challenge for LLMs using zero-shot prompting is their lack of awareness regarding the density of event mention annotations, as precision are much higher than recall. Therefore, we also evaluate LLMs under the one-shot and two-shot settings, and the results are reported in Table 6. Although LLM performance has improved, particularly for GPT-4, whose F1 score significantly increased from 50.7 to 77.7 in zero-shot and two-shot settings, LLMs still perform worse than the supervised baseline.

5.3 Event Coreference Resolution Results

Similar to what we did for event identification, we evaluate LLMs for event coreference resolution under both the zero-shot setting and the few-shot settings. Figure 5 show the results, where we report the averaged F-1 score across the four metrics for event coreference resolution. The detailed results can be found in Appendix A.5. Note that either

	MUC			B^3			$CEAF_e$			BLANC		
	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
GPT-4	6.7	7.3	7.0	92.5	93.3	92.9	86.7	86.7	86.7	4.5	5.0	4.8
Llama-3.1	4.7	20.5	7.6	68.0	94.1	78.9	64.3	64.3	64.3	0.2	15.5	0.5
Mistral-Nemo	1.7	3.3	2.2	85.6	93.1	89.2	80.3	80.3	80.3	0.2	1.7	0.4
QWen-2.5	5.8	10.9	7.6	86.3	93.4	89.7	81.1	81.1	81.1	0.5	7.2	0.9
Supervised Coreference	51.9	64.4	**57.5**	94.2	97.0	**95.6**	95.6	94.0	**94.8**	66.6	84.3	**72.3**

Table 7: Event Coreference Performance. For the LLMs, the evaluation is done in two-shot setting. LLMs significantly underperform the supervised baseline. Supervised baseline has higher performance for B^3, $CEAF_e$, and BLANC since it consider singleton clusters during evaluation, while MUC only consider non-singleton clusters.

one-shot or two-shot prompting do not significantly improve the performance of LLMs, suggesting that simply showing LLMs examples of event coreference relations does not effectively help LLMs better understand the task. The two-shot prompting performs slightly better, yields a small improvement on Mistral-Nemo and achieves comparable performance on other LLMs.

Table 7 shows event coreference resolution results for LLMs and the supervised baseline in the two-shot setting. Scores for B^3 and $CEAF_e$ are much higher because these two metrics consider singleton events during calculation while MUC and BLANC either only consider non-singleton events or have a higher weight on non-singleton events. The supervised approach, leveraging RoBERTa-base fine-tuned on `LegalCore`, significantly outperforms all LLMs across all the evaluation metrics. In contrast, all LLMs struggle with this task, and even the best performing LLM model GPT-4 only obtains very low MUC and BLANC scores. We also noticed that Llama-3.1 has lower B^3 and $CEAF_e$ scores compares to other LLMs. With further investigation, we noticed that Llama-3.1 tends to densely link event mentions and form large noisy coreference clusters that significantly degrade performance for B^3 and $CEAF_e$.

Clearly, all LLMs lag far behind the supervised baseline, revealing the difficulties of LLMs in performing event coreference resolution for legal documents. These results highlight the need for specialized training to bridge the gap between general-purpose LLMs and domain-specific supervised models. We conduct more detailed result analysis in the following paragraphs.

Local vs. Non-local Coreference Clusters Next, we investigate how the models performance differs for local event coreference clusters and non-local event coreference clusters. We report the micro-

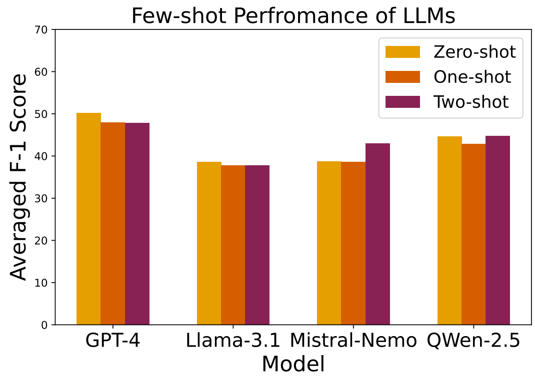

Figure 5: The few-shot performance of LLMs on event coreference. Few-shot prompting doesn't improve LLMs performance.

Model	Local			Non-local		
	P	R	F-1	P	R	F-1
GPT-4	3.0	3.7	3.3	9.2	8.1	8.6
Llama-3.1	1.5	1.5	1.5	2.2	20.6	4.0
Mistral-Nemo	1.5	1.5	1.5	0.7	2.5	1.2
Qwen-2.5	2.7	3.2	2.9	4.1	12.1	6.2
Supervised	43.1	63.9	51.5	44.6	45.8	45.2

Table 8: Event coreference performance for local and non-local coreference clusters. LLMs, comparing to itself, are better at capturing non-local coreference relations than local coreference relations. The supervised model is better at capturing local coreference relations other than non-local relations with longer distance.

averaged MUC scores for local and non-local coreference clusters in Table 8. We only consider MUC scores as MUC is the only metric among all four metrics that only considers non-singleton clusters in calculation, which makes it a better indicator of model performance in this case as singleton clusters are not considered in this analysis.

As shown in Table 8, surprisingly, the performance of LLMs in capturing non-local (cross-section) coreference clusters is higher than their performance on local coreference clusters. One potential reason is that we prompt LLMs to process

Long-distance Dependencies Example
...the Parties {E13 agree} as follows : ... 1.4 " **{E16 Invention}** " means any E17 invention, know-how, data, {E18 discovery} or proprietary information, whether or not patentable, that is made or generated solely by the Representatives of Anixa or OntoChem or jointly by the Representatives of Anixa and OntoChem in performing the Research Plan, including all intellectual property rights in the foregoing. ... 2.7 Records: Each Party will {E134 maintain} complete and accurate records of all {E135 activities} {E136 performed} by or on behalf of such Party under the Research Program and all **{E137 Inventions}** {E138 made} or {E139 generated} by or on behalf of such Party in the {E140 performance} of the Research Program. ...

Table 9: An example of long-distance coreference relations. Coreferential events are denoted in **bold**.

the entire document for identifying coreference relations, therefore, they may overlook local clusters as they prioritize a broader context over specific areas, leading to lower recall in resolving local clusters. In contrast, the supervised baseline is better at resolving local coreference clusters than non-local coreference clusters. However, we believe one main bottleneck of the supervised baseline is its context length limitation. RoBERTa-base has a maximum context length of 512 tokens, which is significantly shorter than the average document length in `LegalCore`. As a result, long documents are split into multiple chunks and encoded separately, making it challenging to identify coreference relations across chunks.

The Challenge in Resolving Long-distance Coreference Relations We observe that both the supervised-learning baseline and LLMs face challenges in resolving coreference relations with two event mentions very far apart. In Table 9, we present an example to spotlight the challenge in resolving long-distance coreference relations. In the example, {E16 Invention} corefers with {E137 Inventions}, despite being 1,264 words apart, spanning two sections and twelve subsections.

Error Analysis We also investigate mistakes made by LLMs and the supervised baseline model. We report the percentage of false positives (FP) and false negatives (FN) among the wrongly predicted coreference links in Table 10.

Notice that for the supervised baseline, the majority of mistakes (63.0%) stem from FP, where the model incorrectly identifies two unrelated event mentions as coreferent. In local coreference links, FP accounts for an even larger proportion (73.3%) of errors. This may be due to the data distribution, as over half of the coreference links in `LegalCore` occur within 50 tokens, leading the supervised model to learn this pattern during training. For non-local coreference links, FP and FN contribute almost equally to the errors.

	Local		Non-local		Overall	
Model	FP%	FN%	FP%	FN%	FP%	FN%
GPT-4	55	45	47	53	52	48
Llama-3.1	49	51	92	8	84	16
Mistral-Nemo	50	50	77	13	67	33
Qwen-2.5	54	46	76	24	67	33
Supervised	73	27	50	50	63	37

Table 10: Rates (%) of different errors for event coreference. False positives (FP) and false negatives (FN) are indicated. Notice that the majority of the mistakes stem from FP for both LLMs and supervised baseline.

For LLMs, false positives (FP) also account for the majority of mistakes, except for GPT-4. Specifically, FP makes up 84.1%, 66.7%, and 66.7% of the overall errors for Llama-3.1, Mistral-Nemo, and QWen-2.5, respectively. However, in contrast to the supervised baseline, false positive mistakes are more common in non-local coeference link prediction. This suggests that LLMs tend to link irrelevant event mentions across sections, and this issue is more pronounced in open-source models. The false positive (FP) rates are especially high for Llama-3.1, upon analyzing the outputs of Llama-3.1, we observed that after initially making some reasonable predictions, very quickly, the model starts to link all the event mentions together, forming a single cluster with numerous event mentions from all over the document.

5.4 End-to-end Results

Table 11 presents the end-to-end results for both LLMs and the supervised baseline. Compared to event coreference results using gold event mentions (Table 7), the coreference resolution performance of LLMs decreased quickly in the end-to-end setting, this is reasonable as LLMs do not perform very well in event identification.

In contrast, the supervised baseline maintains high performance and is much more robust benefiting from explicit training on both event identification and coreference resolution.

	MUC			B^3			$CEAF_e$			BLANC		
	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
GPT-4	1.9	1.7	1.8	60.7	63.1	61.9	70.6	69.3	69.9	0.6	0.7	0.6
Llama-3.1	5.0	21.0	8.0	32.0	48.1	38.1	33.2	47.7	39.0	0.2	13.4	0.4
Mistral-Nemo	1.5	2.8	1.9	35.5	44.5	39.5	53.2	41.8	46.8	0.2	1.3	0.3
QWen-2.5	2.9	5.0	3.7	39.0	47.8	43.0	63.7	41.5	50.3	0.2	2.7	0.4
Supervised End-to-end	50.5	58.3	**54.1**	94.6	96.5	**95.5**	95.3	94.2	**94.8**	67.7	81.3	**72.6**

Table 11: End-to-end Performance. For the LLMs, the evaluation is done in two-shot setting. LLMs significantly underperform the supervised baseline.

6 Conclusion and Future Work

We have presented the first event dataset for the legal domain, `LegalCore`, which has been annotated with comprehensive event and event coreference information. We further benchmark mainstream Large Language Models (LLMs) on this dataset for both event identification and event coreference resolution tasks, and find that this dataset poses significant challenges for both open-source and proprietary LLMs. For future work, we will extend the dataset to cover other event relations in legal documents, such as temporal relations and causal relations.

Limitations

`LegalCore` only contains one type of legal documents, legal contracts, there are many other types of legal documents as well, which can be further considered for event analysis. Another limitation of `LegalCore` is that it only covers coreference relation. The dataset can be more useful if it can be further annotated with other event relations, such as temporal, causal, and subevent relations.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback and input. This research was supported by a gift from Adobe Research at Adobe Inc., and portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

James Allan. 2002. Topic detection and tracking: event-based information organization.

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. The GeneReg corpus for gene expression regulation events — an overview of the corpus and its in-domain and out-of-domain interoperability. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arXiv:2307.03109.

Pei Chen, Kang Liu, Yubo Chen, Taifeng Wang, and Jun Zhao. 2021. Probing into the root: A dataset for reason extraction of structural events from financial documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2042–2048, Online. Association for Computational Linguistics.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. 2022. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3329–3339, Seattle, United States. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2133, Copenhagen, Denmark. 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).

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. *Theory and Applications of Categories*.

Joe Ellis, Jeremy Getman, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2016. Overview of linguistic resources for the tac kbp 2016 evaluations: Methodologies and results. *Theory and Applications of Categories*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie Strassel. 2017. Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. *Theory and Applications of Categories*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 239–249, Sofia, Bulgaria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *Preprint*, arXiv:2103.06268.

Rebecca Hicke and David Mimno. 2024. [Lions: 1] and [Tigers: 2] and [Bears: 3], oh my! literary coreference annotation with LLMs. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Kanae Oda, and Junichi Tsujii. 2007. From text to pathway: Corpus annotation for knowledge acquisition from biomedical literature. In *Asia-Pacific Bioinformatics Conference*.

Xiao Liu, Heyan Huang, and Yue Zhang. 2019a. Open domain event extraction using neural latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871, Florence, Italy. Association for Computational Linguistics.

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

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

Jing Lu and Vincent Ng. 2021b. Conundrums in event coreference resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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).

Mistral AI. 2024. Mistral nemo. Accessed: 2025-02-06.

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

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria. 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, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in ontonotes. In *International Conference on Semantic Computing (ICSC 2007)*, pages 446–453.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50 – 50.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

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

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Knowledge Discovery and Data Mining*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10:349.

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

C. Walker and Linguistic Data Consortium. 2005. *ACE 2005 Multilingual Training Corpus*. LDC corpora. Linguistic Data Consortium.

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

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

Kangda Wei, Aayush Gautam, and Ruihong Huang. 2024. Are LLMs good annotators for discourse-level event relation extraction? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1–19, Miami, Florida, USA. Association for Computational Linguistics.

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

Kangda Wei, Sayan Ghosh, and Shashank Srivastava. 2022. Compositional generalization for kinship prediction through data augmentation. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 13–19, Seattle, United States. Association for Computational Linguistics.

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

A Example Appendix

A.1 Hyper-parameters settings

For the baseline detection model, we train the model with the learning rate sets to $1e-4$ for the T-5 model, and the batch size sets to 4. We train the model for 100 epochs with 5-fold cross validation. For the baseline coreference model, we train the model following the hyperparameters in Wang et al. (2022). We train the model with the learning rate sets to $1e-5$ for the RoBERTa model, the learning rate sets to $1e-5$ for the classification head, and the batch size sets to 4. We train the model for 200 epochs with 5-fold cross validation. All training are conducted on A-100 GPUs.

A.2 LLM Version

We list the detail information of the LLMs evaluated in the paper below:

- **Llama-3.1**: We use the *meta-llama/Llama-3.1-8B-Instruct*[2] from Huggingface [3]
- **Mistral-Nemo**: We use the *mistralai/Mistral-Nemo-Instruct-2407* from Huggingface.
- **QWen-14b**: We use the *Qwen/Qwen2.5-14B-Instruct* from Huggingface.
- **GPT-4-Turbo**: We use the *gpt-4-turbo* accessed through OpenAI API [4]. The experiments are conduct within the window between Jan 15th, 2025 and Feb 15th, 2025.

A.3 Prompts Used

We show the prompts used for benchmarking LLMs foe Event Detection and Event Coreference in Table 12.

A.4 Annotators

All annotators are graduate student from research labs of universities who have great experiences in the field of Natural Language Processing. No additional payments to students are given other than graduate research assistant-ship the students already have.

A.5 LLMs Few-shot Performance

We report the percentage of false positives (FP) and false negatives (FN) of the predicted links in Table 10 where two-shot prompting is used for LLMs. We report the zero-shot and one-shot performance of LLMs on event coreference in Table 14. We report the zero-shot and one-shot performance of LLMs with end-to-end setting in Table 15.

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[3] https://huggingface.co
[4] https://openai.com/api/

Task	Prompt
Event Detection	Please analyze the following text to detect all events. We define an event as any occurrence, action, process or event state which deserves a place upon a timeline, and could have any syntactic realization as verbs, nominalizations, nouns, or even adjectives. Please respond concisely and directly to the point, avoiding unnecessary elaboration or verbosity. If an event is detected, kindly provide its span as index and trigger word/phrase, formatting your response as: Span: event span index Trigger: trigger word/phrase Span: event span index Trigger: trigger word/phrase ... If no event is identified, simply return None. Text: {text} Response:
Event Coreference	Please analyze the following text to detect all coreference relations among events. Two events have a coreference relation if they refer to the same event in space and time. Coreference relation is symmetrical (i.e., non-directional): If A coreferences B, then B coreferences A. It is also transitive: If A coreferences B and B coreferences C, then A coreferences C. Each event is uniquely identified in the text by an identifier in the format {E## trigger_word}. For example: - {E01 discovered} - {E02 collaborated} - {E03 agreed} Response Format: List all coreference relations strictly following this format: E01 COREFERENCE E03 E02 COREFERENCE E05 IMPORTANT: - Use exactly the same event identifiers as in the text. - Do not change the format of the event IDs (always use E##). - If there are multiple coreference relations, list each on a new line. - If no coreference relation is detected, return "None" (do not add any explanation). - If examples are provided, they are for illustration only. Do not copy the event identifiers from the examples. Use only the event identifiers found in the provided text.

Table 12: Prompts used for Benchmarking LLMs for Event Detection and Event Coreference.

	Local		Non-local		Overall	
Model	FP	FN	FP	FN	FP	FN
GPT-4	1116	914	545	623	1633	1509
Llama-3.1	902	935	6103	538	6825	1293
Mistral-Nemo	921	935	2255	661	3154	1574
Qwen-2.5	1075	919	1899	596	2908	1449
Supervised	624	227	530	519	988	580

Table 13: Number of different errors for event coreference. False positives (FP) and false negatives (FN) are indicated.

	MUC			B^3			$CEAF_e$			BLANC		
	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
Zero-shot												
GPT-4	10.9	11.4	11.1	92.9	93.5	93.2	87.3	87.3	87.3	9.5	9.4	9.4
Llama-3.1	6.0	25.4	9.7	68.9	94.5	79.7	64.5	64.5	64.5	0.3	18.9	0.5
Mistral-Nemo	2.2	8.1	3.4	72.8	93.3	81.8	69.0	69.0	69.0	0.4	4.9	0.8
QWen-2.5	4.4	7.4	5.6	87.7	93.3	90.5	81.9	81.9	81.9	0.3	4.1	0.5
One-shot												
GPT-4	8.2	2.3	3.6	98.1	93.1	95.5	91.3	91.3	91.3	4.8	1.1	1.7
Llama-3.1	4.7	21.0	7.7	67.8	94.2	78.8	64.3	64.3	64.3	0.3	14.2	0.5
Mistral-Nemo	4.2	17.0	6.8	70.6	93.9	80.6	66.5	66.5	66.5	0.3	11.5	0.5
QWen-2.5	5.9	16.5	8.7	79.9	93.8	86.3	75.6	75.6	75.6	0.6	10.8	1.1

Table 14: Event Coreference Performance for LLMs with zero-shot and one-shot prompting.

	MUC			B^3			$CEAF_e$			BLANC		
	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
Zero-shot												
GPT-4	1.0	0.4	0.5	32.7	33.3	33.0	67.1	36.3	47.1	0.5	0.1	0.2
Llama-3.1	5.2	19.9	8.2	26.4	44.0	33.0	33.2	35.5	34.3	0.2	14.8	0.5
Mistral-Nemo	2.3	6.9	3.5	30.2	45.9	36.4	46.4	35.5	40.1	0.5	5.0	1.0
QWen-2.5	2.8	4.8	3.5	39.3	48.2	43.3	64.4	42.1	50.9	0.2	2.4	0.4
One-shot												
GPT-4	0.4	0.1	0.2	60.3	58.2	59.2	65.3	74.8	69.8	0.03	0.0	0.1
Llama-3.1	4.2	17.8	6.8	30.2	45.0	36.2	30.3	48.5	37.2	0.3	13.1	0.5
Mistral-Nemo	3.6	13.2	5.7	28.3	49.0	35.8	43.1	31.8	36.6	0.3	8.1	0.5
QWen-2.5	2.4	6.1	3.5	41.1	54.2	46.7	57.7	45.5	50.8	0.2	2.7	0.4

Table 15: End-to-end Performance for LLMs with zero-shot and one-shot prompting.