Okay, Let's Do This! Modeling Event Coreference with Generated Rationales and Knowledge Distillation

Abhijnan Nath, Shadi Manafi, Avyakta Chelle and **Nikhil Krishnaswamy**
Situated Grounding and Natural Language (SIGNAL) Lab
Department of Computer Science
Colorado State University, Fort Collins, CO, USA
{abhijnan.nath, nkrishna}@colostate.edu

Abstract

In NLP, Event Coreference Resolution (ECR) is the task of connecting event clusters that refer to the same underlying real-life event, usually via neural systems. In this work, we investigate using abductive free-text rationales (FTRs) generated by modern autoregressive LLMs as distant supervision of smaller student models for cross-document coreference (CDCR) of events. We implement novel rationale-oriented event clustering and knowledge distillation methods for event coreference scoring that leverage enriched information from the FTRs for improved CDCR without additional annotation or expensive document clustering. Our model using coreference-specific knowledge distillation achieves SOTA B^3 F_1 on the ECB+ and GVC corpora and we establish a new baseline on the AIDA Phase 1 corpus. Our code can be found at https://github.com/csu-signal/llama_cdcr.

1 Introduction

Event Coreference Resolution (ECR) is the task of connecting mentions that refer to the same underlying real-life event. Since descriptions of similar events often use similar words in similar context, systems can achieve strong baseline performance simply by comparing the lemmas of the event triggers (Bugert et al., 2021a; Nath et al., 2023).

However, most ECR datasets contain many event pairs that might be coreferent despite different lemmas, or non-coreferent despite similar lemmas or tokens. Consider this example: "*Video of Brooklyn woman's fatal shooting is played at trial of two men charged in rooftop gunplay.*" The lexical similarity in the event triggers is misleading since they are actually *not* coreferent. Typical heuristic-based systems fail in such cases if the decision is made at the *event pair level*.

Intuitively, a human might solve a challenging coreference problem by engaging in a step-by-step "inner monologue" that reasons about the context, participants, actions, locations, etc., in both events (Bershon, 1992; Alderson-Day and Fernyhough, 2015). Recent works like InstructGPT (Ouyang et al., 2022) have shown that generative large language models (LLMs) can engage in Chain-of-Thought reasoning, a kind of step-by-step reasoning that appears human-like. Such models have also demonstrated abductive reasoning capabilities about relations between events (Zhao et al., 2023b; Ravi et al., 2023) and zero-shot resolution abilities in various coreference benchmarks like CoNLL-2012 and ECB+ (Yang et al., 2022; Le and Ritter, 2023).

In this paper, we seek to model such an inner-monologue or a step-by-step reasoning process in an event coreference system. We augment existing coreference corpora using a novel instruction-based zero-shot prompting framework (Kojima et al., 2023) that guides a generative LLM to produce outputs displaying abductive reasoning about the coreference samples therein.

These intermediate reasoning steps, consolidated into free-text rationales (FTRs) for the coreference labels of the mention pairs, are then used to guide a two-stage modeling procedure. We first perform "Rationale-Oriented Event Clustering (ROEC)" by directly optimizing a "student" model to encode cluster-level information in the coreference graph. Since the generated rationales comes from a disparate model distribution, we simultaneously align event pairs with their corresponding FTRs in the student model's latent space. We then use the optimized student distribution as our backbone encoder, which learns coreference probabilities of event pairs by jointly optimizing the task-supervision component with additional supervision from the generative LLM "teacher" distribution using the rationales as soft labels. Our novel contributions are:

- A method for augmenting coreference datasets with evidence for decisions using FTRs from

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3931–3946

state-of-the-art generative LLMs;

- A novel clustering method to align these rationales with corresponding event pairs;
- A novel optimization framework for distilling contextual cues from these rationales into smaller encoder models using a customized loss function.

Rationales provide an additional soft supervisory signal through which we train the student model with additional information about contextual cues for event coreference, but are not required at inference, increasing our approach's generalizability. We evaluate our method on three event CDCR corpora: Event Coreference Bank Plus (ECB+), the Gun Violence Corpus (GVC), and AIDA Phase 1. Our method achieves state-of-the-art B^3 score on ECB+ and GVC, and sets a novel performance baseline on AIDA Phase 1, without a document clustering step as used in many other methods. We perform detailed ablations of individual components of our method and evaluate how each one contributes to the final performance. Our code, weights, and FTR sets are available at `https://github.com/csu-signal/llama_cdcr`.

2 Related Work

Event CDCR Most previous works in CDCR address the challenge of pairing across documents with a document clustering step that reduces the search space for potential candidates and ensures tractability of pairwise computations (Lee et al., 2012; Yang et al., 2015; Choubey and Huang, 2017; Cattan et al., 2021; Caciularu et al., 2021; Yu et al., 2022). However, preprocessing with document clustering misses a non-trivial amount of coreferring pairs between clusters (Cremisini and Finlayson, 2020). Bugert et al. (2020) also demonstrate that this tends to overfit CDCR systems to corpora like ECB+ (Cybulska and Vossen, 2014), that have unrealistic lexical distinction between topics, thus reducing systems' generalizability to corpora with more referential ambiguity (e.g., GVC; Vossen et al. (2018)). Held et al. (2021) avoid document clustering by modeling discourse focus at the mention level, and Ahmed et al. (2023) leverage heuristics that capture discourse-level lemmatic features.

While Held et al. (2021) also suggest using knowledge distillation techniques (Gou et al., 2021) to enhance pairwise computations even further, using a heuristic allows us to approach knowledge distillation using a single encoder to model both cluster-level and pairwise signals without using separate encoders for candidate retrieval and candidate scoring. In line with this, we use Ahmed et al. (2023)'s heuristic-based approach for candidate generation during training and inference.

Abductive Reasoning and Free-Text Rationales Previous work on abductive reasoning in coreference resolution has used rule-based and entity-specific approaches, with causal relations between events only appearing as supporting evidence for the coreference decision (Raina et al., 2005; Inoue et al., 2012; Yamamoto et al., 2015). Zhao et al. (2023b) and Bhagavatula et al. (2019) explore Bayesian methods to model the plausibility of abductive explanations and suggest mutual exclusivity of such explanations. The rise of autoregressive general-purpose LLMs have inspired works like Ho et al. (2022), Snell et al. (2022), Shridhar et al. (2023), Narang et al. (2020), Sun et al. (2022), and Rajani et al. (2019), which use LLM-generated FTRs for NLI tasks like commonsense reasoning, for both training and inference. Wiegreffe et al. (2021) and West et al. (2022) explore knowledge distillation techniques using such FTRs and suggest evaluation frameworks to assess their quality. Ahmed et al. (2024a,b) explores the capabilities of LLMs like GPT-4 in generating argumental and temporal information between entities in order to augment the annotation process in event coreference corpora like ECB+.

Research in cognitive psychology (Alderson-Day and Fernyhough, 2015) suggests that a well-developed "inner monologue" is crucial for different aspects of problem-solving, often as a "working memory" or a cognitive rehearsal tool (Sokolov, 2012). Ravi et al. (2023) use this strategy for temporal reasoning in GPT-3-generated FTRs to improve coreference computations. However, due to the black-box nature of LLMs like GPT-3, their method assumes the FTRs themselves as the complete knowledge, with no access to the LLM's internal distribution. In contrast, our research uses an open-weight LLM both to generate step-by-step FTRs with reasoning about coreference, and so that we can access the underlying model distribution.

3 Method

Our method for event coreference resolution consists of three parts: 1) Using curated zero-shot instructions specifying the ECR task, we gener-

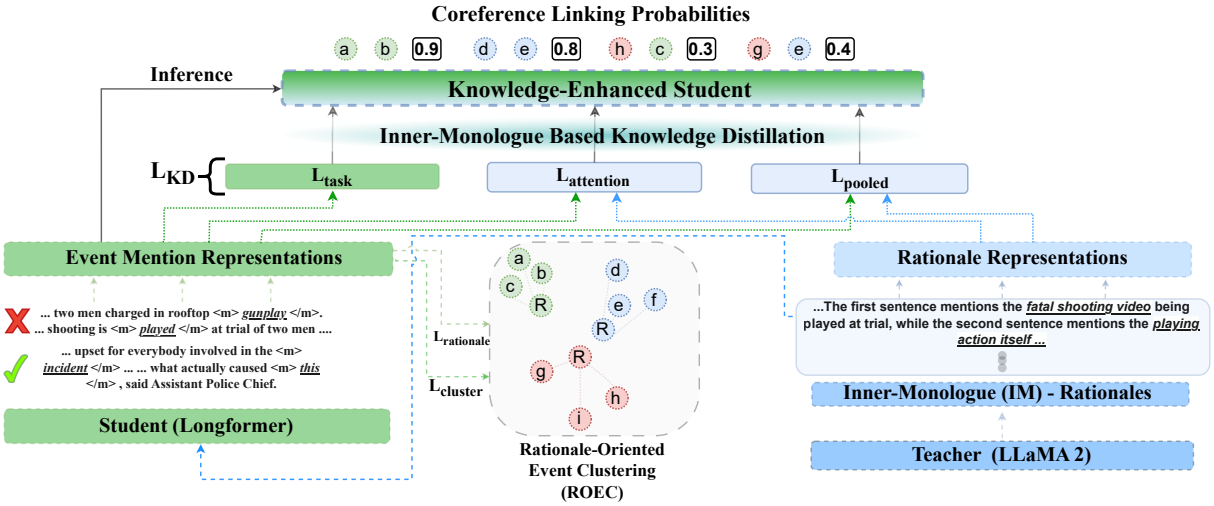


Figure 1: Schematic system overview: Step-by-step FTRs resembling an "inner-monologue" are generated using an LLM (teacher model) conditioned on the gold coreference label. FTRs are then clustered along with event pairs to optimize the student's latent space (ROEC). The optimized student learns further coreference-specific contextual cues in the rationales from the teacher's latent space. Arrows show the gradient flow during training from the teacher (blue) and student (green) during the ROEC (dashed) and knowledge distillation (dotted) phases, respectively. Solid black line indicates inference samples, which include no rationale text or signal from the teacher model. Letters a-i in the ROEC block represent distinct event mentions, and the colors represent an event cluster (such that all the blue circles cluster together). "R" represents a set of rationales that justify the linking of different mentions in a single cluster.

ate abductive rationales from a generative LLM teacher model; 2) We implement a Rationale-Oriented Event Clustering (ROEC) procedure that calibrates the student model distribution with gold-standard event clusters while drawing distant supervision from the previously-generated event-specific rationales; 3) We train the student model along with a frozen teacher distribution to model coreference probabilities of sampled event pairs.

Fig. 1 provides a schematic of our approach. In this paper, we use LLaMA 2-7B-Chat (Touvron et al., 2023) for the teacher model and Longformer-base (Beltagy et al., 2020) for the student model.

3.1 Datasets

We evaluate our method across three English event CDCR corpora with varying levels of referential ambiguity and difficulty.

Event Coreference Bank Plus (ECB+) Most prior works in event CDCR have evaluated on ECB+ (Cybulska and Vossen, 2014) due to the wide variety of topics that it covers. The lexica used in different ECB+ topics are largely distinct, leading to less overall ambiguity (Bugert et al., 2020). We follow Cybulska and Vossen (2015)'s approach for training, validation and testing splits.

Gun Violence Corpus (GVC) The GVC (Vossen et al., 2018) contains annotated events specifically in the domain of gun violence. The similar lexicon across event mentions leads to a high referential ambiguity. This tends toward coreference chains with a more realistic (i.e., non-Zipfian) distribution of events in text descriptions. This makes GVC more challenging especially in a CDCR setting. We use the splits from Bugert et al. (2021a).

AIDA Phase 1 The AIDA Phase 1 corpus (Tracey et al., 2022) consists specifically of events in the domain of the Russia-Ukraine conflict, which are annotated based on their potential for conflicting perspectives.[1] This corpus's test set is larger than its training set, which makes it additionally challenging. To the best of our knowledge, the only evaluation to date performed on this dataset was performed by members of our team (Nath et al., 2024). As before, we follow the splits from Tracey et al. (2022).

3.2 Event Coreference Rationale Generation

We define rationale generation with LLMs as an abductive reasoning problem (Paul, 1993). Given a pair of event mentions and their contexts as an

[1]The data is available from the Linguistic Data Consortium under catalog number LDC2019E77.

observation (e_1 and e_2) and the coreference gold label as an *outcome* (g), the LLM should generate the most probable *hypothesis* or *rationale* (r^*), where:

$$r^* = \arg\max_{r^i} P\left(R = r^i \mid e_1, e_2, g\right) \quad (1)$$

We assume mutual exclusivity of rationales (Gordon and Hobbs, 2017), such that one plausible rationale automatically excludes other rationales for an event mention pair. This allows us to rewrite Equation 1 using Bayes Rule conditioned on the gold standard (g) as:

$$P\left(R \mid e_1, e_2, g\right) \propto P\left(g \mid e_1, e_2, R\right) \cdot P\left(R \mid e_1, e_2\right) \quad (2)$$

Since the gold coreference labels are meant to be the ground truth, we have:

$$P\left(R \mid e_1, e_2, g\right) \propto P\left(R \mid e_1, e_2\right) \quad (3)$$

We use gold coreference labels in our prompts to guide rationale generation. This establishes a one-to-one map[2] between an event mention pair and its rationale while also reducing the cost (computational or otherwise) associated with additional hypothesis generation.

Rationales are generated with LLaMA 2-7B-Chat (Touvron et al., 2023), a foundation model fine-tuned for "human-like" conversation. LLaMA 2's open-weight nature means the model distribution remains accessible. Since the model is otherwise not fine-tuned for event coreference, we prompt the language model to ground its rationale to event coreference-specific arguments such as participants, times, entities, and locations. This provides event-specific context that grounds the output to the event mention pair. This assures mutual exclusivity and provides information that can be used in representational learning techniques (Murty et al., 2020; Kenyon-Dean et al., 2018) for aligning or "interpreting" an event mention within the context of a rationale. For sampling, we use a temperature parameter of 0.7 for randomness and a top-p of 0.9 to ensure diversity in the tokens. We constrain generation to a maximum of 512 tokens. See Fig. 2 for prompt formatting, and Table 1 for our rationale dataset statistics. These rationales or intermediate reasoning steps are then used to guide the other two stages of our procedure (Secs. 3.4 & 3.5).

[2]Coreference annotation typically relies on a structured knowledge base to minimize this set of plausible hypotheses (Vossen et al., 2018; Tracey et al., 2022).

Task Instructions (System Prompt):

You are a coreference annotator.
Think step by step and create an inner monologue to answer a question about sentences A and B.
Two events are marked with **<m>** and **</m>** tokens. Identify common context, actions, participants, objects and locations.
Come to a conclusion whether the two events are the same based on your inner-monologue

Abductive User Prompt

Sentence A: The actress **<m>** **rehired** **</m>** longtime attorney Shawn Holley to handle her case.
Sentence B: Lindsay Lohan **<m>** **rehires** **</m>** longtime lawyer Shawn Holley , heads to rehab.
Answer: These two events are coreferent.
Your inner-monologue:

Figure 2: Prompt format for inner monologue-based FTR generation conditioned on the gold label (underlined). <m> and </m> demarcate the event triggers.

3.3 Descriptive Statistics of Generated Rationales

	Corpus		
	ECB+	GVC	AIDA
# Event Pairs	41,334	80,060	17,306
# Total Tokens	12.3M	24.5M	5.3M
# Unique Tokens	12.2k	13.0k	7.5k
Avg. Token Length	4.7	4.6	4.8
Avg. Tokens/FTR	300	305	310
Self-BLEU	.77 (.66)	.82 (.75)	.79 (.78)

Table 1: Descriptive statistics of generated inner monologue-based free text rationales (FTRs) for ECB+, GVC, and AIDA Phase 1 at the corpus level. Self-BLEU scores of the gold coreference mentions are shown within parentheses.

In order to assess the diversity and the uniqueness of the generated rationales, we also conducted a lexical diversity analysis at the token and n-gram level. LLM-generated explanations typically tend to be less diverse and unique compared to human-written examples (Welleck et al., 2019; West et al., 2022). We did not collect a large sample of human-generated rationales to compare to, but while a token-level diversity can provide an estimate of the exact lexical distribution in the rationales, a "softer" estimation can be drawn using the Self-BLEU (Zhu et al., 2018) metric. This gives us an idea of how similar rationales are amongst themselves at the n-gram level (higher value means more similar).

We report Self-BLEU scores for generated rationales for each corpus using a 10% random sampling with a fixed seed. Table 1 shows that lexical uniqueness (unique tokens) is likely a function of the overall lexical distinctness of the corpus

and not directly proportional to the cardinality of the full token set. For instance, ECB+ has almost twice the proportion of unique tokens to total token count when compared to GVC. Additionally, Self-BLEU scores of the corresponding gold coreference mentions (drawn directly from the corpora) are generally lower than those of FTR sets. More soft-uniqueness or n-gram diversity in the gold mentions could be due to the step-by-step nature of abductive reasoning where multiple angles of reasoning, and therefore wording, can co-occur with similar context. We also conducted a human evaluation of the FTRs across various markers of fidelity and quality motivated by (Wiegreffe et al., 2022). Appendix E contains details of the human evaluation component.

3.4 Rationale Oriented Event Clustering (ROEC)

We perform "Rationale-Oriented Event Clustering" (ROEC) by directly optimizing the student model to encode cluster-level information in the coreference graph. Since the generated rationales come from a disparate model distribution, we align event pairs with their corresponding rationales in the student model's latent space.

Positive training samples consist of all mention pairs belonging to the same coreference cluster. When collecting negative samples, we want to avoid overwhelming the training distribution with non-coreferent pairs, which constitute most of any CDCR corpus. Since a rationale is assumed to be a supporting hypothesis for a specific event mention pair, as opposed to any other pair or at the cluster level, minimizing the search space is crucial, especially without a document-clustering step. As such, we sample negative pairs using the non-oracle heuristic from Ahmed et al. (2023) with a low-threshold of 0.05. This heuristic, depending on the threshold, selectively retrieves hard negative pairs using a cache of synonymous lemma pairs occurring in the coreference chains. Since ECB+ and GVC contain almost no inter-cluster coreferents (Held et al., 2021), this makes the process of generating negative samples tractable and avoids non-informative inter-cluster candidates.

We encode an event mention pair ($e1$, $e2$) into a common representation $\overrightarrow{p}$ by extracting the `[CLS]` token representation of the concatenated input from the last hidden layer of the student model. We encode the supporting rationale $\overrightarrow{r}$ similarly in the student. For training, we use a joint-optimization framework: a cross-entropy based cluster loss ($\mathcal{L}_{\text{cluster}}$) (Eq. 4) to predict the cluster label of the event pairs, and a cosine-distance based loss ($\mathcal{L}_{\text{rationale}}$) (Eq. 5) with a tuned weight penalty (λ) to align the vector embeddings for the rationales to those of the corresponding event pairs. For training batch size m, total number of clusters[3] N, predicted cluster probabilities $\hat{y}_i$, and ground truth cluster vector y_i, we minimize a joint-optimization loss given by Eq. 6, with components (Eq. 4) and (Eq. 5):

$$\mathcal{L}_{\text{cluster}} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{n=1}^{N+1} y_i \cdot \log(\hat{y}_i) \tag{4}$$

$$\mathcal{L}_{\text{rationale}}\,(\overrightarrow{p}\,,\overrightarrow{r}) = \sum_{i=1}^{m}\left(1 - \frac{\overrightarrow{p_i}\cdot\overrightarrow{r_i}}{\|\overrightarrow{p_i}\|\,\|\overrightarrow{r_i}\|}\right) \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{\text{cluster}} + \lambda\mathcal{L}_{\text{rationale}} \tag{6}$$

3.5 Coreference Knowledge Distillation

We use the optimized student model from the previous step as the backbone encoder. Our classifier then learns coreference probabilities of event pairs by optimizing the task-supervision component against gold labels while simultaneously aligning the student distribution with the teacher model representations of the rationales as soft labels. We optimize a "knowledge distillation loss" given by Eq. 7.

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{\text{attention}} + \lambda_2\mathcal{L}_{\text{pooled}} \tag{7}$$

We estimate the regularization parameters with a grid search over the validation set. We find λ_1 = 1 and λ_2 = 0.01 to work best for student model convergence. Each individual component is defined as follows.

Task Component This uses a pairwise scorer framework (Caciularu et al., 2021) to train the student model. Document pairs containing the individual event-trigger spans are encoded in the student model into a common representation that consists of the `[CLS]` representation of the document pair and the attention components of representations e_1 and e_2 as well as of Hadamard product $e_1 \odot e_2$.

[3]Following Kenyon-Dean et al. (2018), we cluster all singletons into a single dummy label. Following Rahman and Ng (2009), negative pairs are assigned to this $N + 1^{\text{th}}$ class in $\mathcal{L}_{\text{cluster}}$ (Eq. 4).

This common representation is then fed into the classification layer (multi-layer perceptron) of the student model to estimate coreference probabilities. For this task-specific supervision, we minimize the binary cross-entropy loss, $\mathcal{L}_{\text{task}}$:

$$\mathcal{L}_{\text{task}} = -\frac{1}{m}\sum_{i=1}^{m}(y_i \cdot \log \hat{y}_i + (1-y_i)\cdot \log(1-\hat{y}_i)) \tag{8}$$

where y and $\hat{y}$ are the true and predicted coreference probabilities in a sample batch of size m.

Distillation with Rationales For the distant supervision-based distillation, aligning the student (S) and teacher (T) distributions is carried out with the rationales ($\boldsymbol{R}$) encoded across the attention states ($\boldsymbol{R}^*_a$) and the last hidden states ($\boldsymbol{R}^*_h$).

Attention Loss Transformer-based language models like BERT tend to capture high-level linguistic knowledge, including coreference signals, in their attention states, distributed across the various heads (Clark et al., 2019). To align the attention states, we minimize the squared L^2 norm between the final-layer attention representations of the rationales as encoded in the student ($\boldsymbol{R}^S_a$) and teacher ($\boldsymbol{R}^T_a$). Motivated by Jiao et al. (2020), we apply a mapping function ($f(i) = i + H - h; \quad 0 < i \leq h$) from student attention head indices to teacher heads, where the i^{th} student head sources supervision from the $f(i)^{th}$ teacher head. H and h represent the total number of attention heads in the teacher and the student, respectively. For instance, the first student attention head is mapped to teacher attention head (1 + H - h). Therefore,

$$\mathcal{L}_{\text{attention}} = \sum_{i=1}^{h} \|\boldsymbol{R}^S_{a_i} - \boldsymbol{R}^T_{a_{f(i)}}\|_2^2 \tag{9}$$

Hidden-state Loss Similarly, here we extract the final-layer pooled rationale representations from the student and teacher, and minimize the squared L^2 norm between them. A learnable linear projection matrix $\boldsymbol{W}_{T\rightarrow S}$ is used to project the teacher's 4096D hidden representation into the student's 768D latent space, resulting in:

$$\mathcal{L}_{\text{pooled}} = \|\boldsymbol{R}^S_h - \boldsymbol{R}^T_h \boldsymbol{W}_{T\rightarrow S}\|_2^2 \tag{10}$$

4 Experiments

Ablations We evaluate four different variants of our model, to establish how each component contributes to final performance. 1) $Long_{\text{paired}}$ establishes a baseline using no ROEC or knowledge distillation. This resembles traditional representational learning systems that leverage natural language rationales. We implement Murty et al. (2020)'s method and "pair" rationales with the corresponding event mentions by extracting the [CLS] token representation from the pairwise scorer framework which is trained using a simple BCE loss. 2) $Long_{\text{+ROEC,-KD}}$ includes ROEC optimization with additional training with a task-specific BCE loss ($\mathcal{L}_{\text{task}}$). 3) $Long_{\text{-ROEC,+KD}}$ excludes ROEC but includes coreference knowledge distillation. 4) $Long_{\text{+ROEC,+KD}}$ uses both components.

Training Parameters For training the ROEC phase (Sec. 3.4), we use an Adam optimizer (Kingma and Ba, 2014) with a batch size of 40 and a model learning rate of $1e-5$ for 20 epochs. For training coreference knowledge distillation (Sec. 3.5) we use a smaller batch size of 16 to ensure optimal performance. We use a model learning rate of $1e-5$ and a classifier learning rate of $1e-3$ and train for 10 epochs. We use a single NVIDIA A100 GPU for training both phases. ROEC and coreference knowledge distillation take roughly 20 minutes and 45 minutes, respectively, for a training a single epoch.

Inference and Evaluation We follow a simple connected components-based clustering approach at inference to generate coreference chains. *No FTRs are included in the input to the model at inference time.* We score candidate pairs with only the gold coreference labels using our coreference knowledge-enhanced model (Sec. 3.5). These scores are then used to construct an affinity graph to identify connected components using a threshold of 0.5. Thereafter, the generated coreference chains are evaluated against the gold clusters to calculate final coreference metrics. Following Held et al. (2021), we focus on the B^3 metric, which is sensitive to incorrectly clustered singletons.

5 Results

Table 2 shows evaluation results over the ECB+, GVC, and AIDA test sets. Results provided are single runs after robust hyperparameter tuning on the validation sets. We compare our method (with ablations) to relevant previous baselines, focusing on the B^3 metric as mentioned above. Since we preprocess according to Ahmed et al. (2023)'s heuristic that samples from a cache of coreferent lemma

pairs built only over the training set (see Sec. 2), we compare to their results reported using this heuristic. Appendix C shows results according to other common coreference metrics.

To demonstrate that a generative LLM alone does not perform equivalently, we compare to zero-shot results from LLaMA 2-7B-Chat (our teacher model) and GPT-3.5-Turbo. These models were prompted to provide a single-word (*yes*/*no*) answer to whether or not the events given in the prompt are coreferent. Prompt format is a slight variant of that given for rationale generation in Fig. 2. See Appendix D for more details.

	B^3		
Methods	ECB+	GVC	AIDA
Bugert et al. (2021a)	-	59.4	-
Cattan et al. (2021)	81.0	-	-
Caciularu et al. (2021)	85.6	-	-
Held et al. (2021)	85.7	83.7	-
Ahmed et al. (2023)	82.4	77.7	-
LLaMA 2-7B-Chat	77.7	53.5	47.8
GPT-3.5-Turbo	79.8	49.6	56.0
$Long_{\text{paired}}$	81.8	75.3	58.1
$Long_{\text{+ROEC,-KD}}$	85.9	80.6	61.2
$Long_{\text{-ROEC,+KD}}$	84.4	82.5	61.5
$Long_{\text{+ROEC,+KD}}$	**86.8**	**84.3**	**64.5**

Table 2: B^3 F_1 results using our coreference knowledge distillation framework on the ECB+, GVC and the AIDA Phase 1 test sets.

Our best model, using both ROEC and knowledge distillation, outperforms previous baselines on both ECB+ (+1.1 B^3 F_1) and GVC (+0.6 B^3 F_1), as well as zero-shot LLaMA 2-7B-Chat and GPT-3.5-Turbo. AIDA Phase 1 is a new, challenging dataset, where we establish a new baseline.

$Long_{\text{+ROEC,+KD}}$ learns from the teacher distribution, but we find it significantly outperforms both the zero-shot teacher model (LLaMA 2-7B-Chat) and the much larger GPT-3.5-Turbo. Without task-specific finetuning, general purpose models tend to under-perform in reasoning-based tasks compared to smaller fine-tuned variants (Ho et al., 2022), but our results suggest that the teacher-generated rationales contain useful information for coreference decisions, such that when they are conditioned on the gold label, the distilled knowledge optimizes the task-based encoder for better performance, despite such larger models performing poorly in a zero-shot setting (Wiegreffe et al., 2022).

5.1 Ablation Tests

When we ablate the knowledge distillation component, we find that adding KD to a model that only performs ROEC ($Long_{\text{+ROEC,-KD}}$) boosts performance by +0.9 B^3 F_1 (ECB+), +3.7 B^3 F_1 (GVC) and +3.3 B^3 F_1 (AIDA). This further attests to the informativeness of the generated FTRs.

Performing ROEC alone results in a 3–5 B^3 F_1 performance boost compared to the simple paired representation learning approach ($Long_{\text{paired}}$). Adding KD on top of that boosts performance more on the GVC and AIDA corpora than on ECB+. This suggests that the knowledge distillation component that sources task-specific supervision from the teacher brings an additional performance gain when there is more referential ambiguity, as in GVC, or conflicting event descriptions, as in AIDA Phase 1. In ECB+, topics are lexically distinct which results in rationales for a given topic having a similar lexical distribution. This suggests that when ROEC alone boosts performance more than KD alone, it is likely due to the nature of the corpus. This is consistent with Bugert et al. (2021a)'s observation that CDCR systems with document clustering tend to overfit to the structure and link distribution of ECB+ largely due to its lexical distinctions between topics.

6 Discussion

We ran Ahmed et al. (2023)'s pipeline over the same data and compared where their method and ours make different decisions on ECB+ and GVC samples. Since we use the same preprocessing here, this compares the performance of the two pairwise discriminators used: plain Longformer vs. our knowledge-enhanced version.

	ECB+		GVC	
	$Long_{\text{+ROEC,+KD}}$	$Long$	$Long_{\text{+ROEC,+KD}}$	$Long$
# pos.	825 (3506)	11 (3506)	470 (2633)	20 (2633)
# neg.	483 (2041)	334 (2041)	428 (6768)	70 (6768)

Table 3: Number of positive (coreferent) and negative samples, per dataset, on which the indicated model succeeded and the other failed. Total number of pairs of the given label in each dataset is shown in parentheses.

Table 3 shows the comparative error analysis. $Long_{\text{+ROEC,+KD}}$ succeeds at linking coreferent ECB+ pairs that plain Longformer fails on 75x as often as the reverse (825 to 11). The margin is 23.5x on GVC. While the margin is lower on non-coreferring samples, $Long_{\text{+ROEC,+KD}}$ is still a substantially stronger performer than plain Longformer here as well.

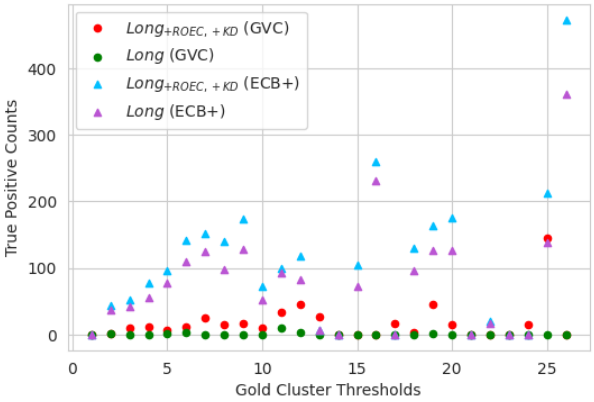


Figure 3: Distribution of mentions correctly resolved by the indicated model vs. cluster size.

Fig. 3 shows the distribution of mentions correctly resolved by the indicated model as a function of cluster size in the two corpora. $Long_{\text{+ROEC,+KD}}$ is nearly globally more successful at making coreference links, and disparities grow as cluster size increases. This effect is seen in both corpora but is particularly pronounced in ECB+.

Using knowledge distilled from FTRs, we are able to make more correct coreference decisions than plain Longformer, even without using the rationale in the input at inference time. This holds true across both ECB+ and GVC test sets. Our method also establishes a new baseline for the AIDA Phase 1 corpus.

Qualitative Analysis Table 4 presents two test samples from each dataset that our best model clustered successfully and plain Longformer did not. We present snippets of the FTRs generated for these samples as well. As FTRs were not included in the input to the model during inference, these simply serve to provide an additional way of interpreting the results, by showcasing the kind of information contained within an FTR that, when distilled into the student model, plausibly contributes to a correct coreference linkage:

(a) Although both documents refer to a 6.1 magnitude earthquake in Aceh, the mention trigger in Document B actually refers to the 2004 tsunami that hit the same region. The FTR mentions this distinction in context.

(b) The two mentions contain the same elements (organizations, prices) but in a different order. The FTR correctly ascribes the organizations and the $5.4 million price to both documents.

(c) The two events take place in the same city and the mentions actually have identical syntax. However, the FTR correctly identifies that they refer to different people and regions/neighborhoods of Baltimore.

(d) Only Document A mentions the location while only Document B mentions the name of the wounded/deceased, but the FTR is able to note that in one document there were 5 victims while in the other there were 1+4 victims.

This shows that in areas like exposing alignment or divergence between named entities, temporal contexts, or syntax across mention pairs, FTRs are providing useful information by making explicit what may be implicit in the raw text. Our knowledge distillation procedure is able to take this information from the attention heads and hidden states of the teacher model and project it into the student model for finer-grained coreference decisions.

7 Conclusion and Future Work

In this paper, we presented a novel event CDCR technique that used free-text rationales from a generative LLM to provide additional supervisory signals for coreference. We accomplished this through a combination of clustering rationale representations with the corresponding event mentions in a student model's space, and by distilling information from the generative teacher model into the smaller student. We achieved SOTA B^3 scores on the ECB+ and GVC benchmarks, and established novel benchmark performance on the challenging AIDA Phase 1 dataset. We also ablated the contributions of different components of our model and examined the kinds of information our model is likely leveraging toward its performance.

It would be too strong a claim to say that the generated FTRs consistently display coherent reasoning in and of themselves (for example, sample (b) in Table 4 seems to treat the two mentions as events occurring in sequence rather than to be considered in parallel; this may be an artifact of the autoregressive generation mechanism), but by forcing the LLM to explicitly generate output concerning the common entities and arguments of the events in question, it is also forced to generate information more explicitly relevant to coreference links than just the event triggers themselves and immediately surrounding context. This resembles a "dense paraphrasing"-like procedure (Tu et al., 2023) that

	True Label	Document A	Document B
(a)	*Not coreferent*	6.1-magnitude quake strikes Indonesia's Aceh province	6.1-magnitude quake strikes Indonesia's Aceh province in region hit by 2004 tsunami
	FTR sample: *"In Document A, the earthquake is the primary focus, while in Document B, the earthquake is mentioned in the context of a tsunami that occurred in 2004."*		
(b)	*Coreferent*	Advanced Micro Devices (NYSE: AMD) announced the largest acquisition in its history Monday, paying about $5.4 billion to acquire graphics chip specialist ATI Technologies.	RBC Capital Markets downgraded ATI Technologies to "sector perform" from "outperform" after Advanced Micro Devices announced its intention to buy the Canadian graphics chip-maker for $5.4 billion on Monday.
	FTR sample: *"[...] I see the phrase 'paying about $5.4 billion to acquire graphics chip specialist ATI Technologies...' Oh, so it's ATI Technologies that AMD acquired! Fast forward to Document B, and I see the same phrase 'Advanced Micro Devices <m> announced </m> its intention to buy the Canadian graphics chip-maker for $5.4 billion on Monday...' Wait, what?! Didn't we just establish that AMD acquired ATI Technologies?"*		
(c)	*Not coreferent*	D'kai Vanlandingham, 17, of Northeast Baltimore, was shot and killed just after 10 p.m. Saturday on the 3400 block of Juneway in Belair-Edison, police said.	Martin Ibarra, 35, of South Baltimore, was shot and killed shortly before 10 p.m. Friday on the 1400 block of Ramsay St. in New Southwest-Mount Clare, police said.
	FTR sample: *"In document A, the victim is identified as D'kai Vanlandingham, while in document B, the victim is named Martin Ibarra. [...] Document A states that the incident happened on the 3400 block of Juneway in Belair-Edison, while document B mentions the 1400 block of Ramsay St. in New Southwest-Mount Clare. [...] [T]here are two separate incidents involving people being shot and killed in different locations in Baltimore."*		
(d)	*Coreferent*	Two suspects were arrested Monday in the killing of a 38-year-old man who died more than two weeks after he and four others were wounded in a shooting outside a Little Rock home.	Five people, including Duhart, were injured. The other victims involved in the shooting were identified as:
	FTR sample: *"[...] Document A is clear and specific [...]: The event occurred outside a Little Rock home. One person was killed, and four others were injured. The event involved a shooting. From Document B, I know that five people, including Duhart, were injured [...]."*		

Table 4: Test samples from ECB+ (a–b) and GVC (c–d) that our method ($Long_{+\text{ROEC},+\text{KD}}$) links correctly and plain Longformer (reproduced according to Ahmed et al. (2023)) fails on. Mention triggers are underlined. As these are test samples, the presented abridged FTR samples are not included in the input for inference, but are given as examples of the kinds of information our model is likely to incorporate in making its decision.

provides enriched event descriptions that we then distill into our model using ROEC and knowledge distillation from the teacher distribution.

Our results demonstrate that although imperfect, automatically-generated rationales for event coreference contain useful information toward the decision. Using AI-generated rationales as soft-labels might prove useful as a way to decrease annotator workload in cognitively-heavy tasks like annotating coreference resolution corpora (Zhao et al., 2023a).

This opens the way to future work to improve the utility of FTRs. For example, filtering methods may be used to exclude FTRs with lower-quality coreference knowledge. Techniques such as West et al. (2022) that employ a separate but smaller critic model can be trained on a small sub-sample of high-quality rationales written by trained coreference annotators, to further enhance coreference-specific knowledge distillation with lower annotation expenses. FTRs generated using a more powerful model like GPT-4 could also be beneficial in extending extant CDCR corpora with more explicit soft-labels and can likely enhance systems that need to detect cross-subtopic coreference in corpora such as FCC (Bugert et al., 2020), albeit at the cost of accessibility to the source model. FTRs with validated gold cluster-level information could be leveraged especially in preclustering to reduce cross-computations, making such systems more generalizable.

Limitations

While our results demonstrate that automatically-generated rationales for event coreference contain useful information toward the decision, there remains the fact that like all current generative AI models, LLaMA 2-7B-Chat may "hallucinate" or output fallacious information. For instance, for a GVC mention pair A: "*3-year-old shot, killed in Stockon while riding in car.*" and B: "*The girl, identified as Melanie Martinez of Stockton, was the only person in the vehicle who was hit by the shots, and her family drove her straight to a nearby hospital, according to police. She was pronounced dead at the hospital.*", the generated FTR mentions that "*Both documents mention the girl was pronounced dead at a hospital.*" However, this fact was only mentioned in the second document. We designed our study with the goal of minimizing such occurrences (see Appendix B), but the frequency and effect of such hallucinations remains to be investigated.

On the zero-shot results comparisons, there may be some sensitivity to the prompt given to the two competing models. Some surface-level prompt engineering was conducted to ensure that the models only provided one word answers in the zero-shot setting and could make a coreference distinction in clear cases, in order to provide a reasonable and evaluable baseline comparison. As the focus of this paper is not on prompt engineering for closed

models, we did not investigate this further.

Our results should be considered in the context of the entire pipeline. There are many ways to preprocess CDCR corpora to render the task more tractable. We eschewed the document clustering step of many popular methods due to computational expense and tendency to exclude many valid inter-cluster links (see Sec. 2). That left us two filtering strategies from recent work: Held et al. (2021)'s discourse modeling in the latent space and Ahmed et al. (2023)'s heuristic filtering. Ahmed et al. (2023)'s method is faster, so we used this preprocessing step. Despite this, we were still able to exceed Held et al. (2021)'s B^3 F_1 scores on ECB+ and GVC.

Given the nature of their subject material, the Gun Violence and AIDA Phase 1 corpora may be troubling to some, including, apparently, a generative LLM. For 167 of 7,314 (∼2.28%) of AIDA Phase 1 samples and 80 out of 10,355 (∼0.77%) of GVC, LLaMA 2-7B-Chat would not generate a definite answer for the event pair when evaluated in the zero shot setting, citing ethical and moral standards and the fact that the event mentions (and therefore prompt) contained descriptions of violence and harm. These samples had to be discarded from evaluation. This effect was not observed when using LLaMA 2-7B-Chat to generate the FTRs for training our methods.

Acknowledgments

Okay, let's do this! We'd like to thank the evaluators of our free-text rationale examples: Nada Alalyani, Jack Fitzgerald, Rahul Ghosh, Anju Gopinath, Huma Jamil, Changsoo Jung, Ibrahim Khebour, and Hannah VanderHoeven. Our thanks also go out to the anonymous reviewers whose comments helped improve the final version of this paper. This material is based in part upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112490377. Views expressed herein do not reflect the policy or position of the Department of Defense or the U.S. Government. All errors are the responsibility of the authors.

References

Shafiuddin Rehan Ahmed, George Baker, Evi Judge, Michael Regan, Kristin Wright-Bettner, Martha Palmer, and James H. Martin. 2024a. Linear cross-document event coreference resolution with x-amr. In *Proceedings of the Joint Conference of the 15th Language Resources and Evaluation Conference, and, the 30th International Conference on Computational Linguistics*, Torino, Italy. European Language Resources Association.

Shafiuddin Rehan Ahmed, Jon Cai, Martha Palmer, and James H. Martin. 2024b. X-amr annotation tool. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Malta. Association for Computational Linguistics.

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

Ben Alderson-Day and Charles Fernyhough. 2015. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141(5):931.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv e-prints*, pages arXiv–2004.

Barbara L Bershon. 1992. Cooperative problem solving: A link to inner speech. *Interaction in cooperative groups. The theoretical anatomy of group learning*, pages 36–48.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story @ ECIR*, pages 23–29.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021a. Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora. *Computational Linguistics*, 47(3):575–614.

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

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E 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.

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

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Andres Cremisini and Mark Finlayson. 2020. New insights into cross-document event coreference: Systematic comparison and a simplified approach. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 1–10, Online. 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.

Agata Cybulska and Piek Vossen. 2015. Translating granularity of event slots into features for event coreference resolution. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Andrew S Gordon and Jerry R Hobbs. 2017. *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

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

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Naoya Inoue, Ekaterina Ovchinnikova, Kentaro Inui, and Jerry R Hobbs. 2012. Coreference resolution with ilp-based weighted abduction. In *Proceedings of COLING 2012*, pages 1291–1308.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. *arXiv preprint arXiv:1805.10985*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Nghia T Le and Alan Ritter. 2023. Are large language models robust zero-shot coreference resolvers? *arXiv e-prints*, pages arXiv–2305.

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

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

Shikhar Murty, Pang Wei Koh, and Percy Liang. 2020. ExpBERT: Representation engineering with natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2106–2113, Online. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Abhijnan Nath, Huma Jamil, Shafiuddin Rehan Ahmed, George Baker, James H. Martin, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Multimodal cross-document event coreference resolution using linear semantic transfer and mixed-modality ensembles. In

Proceedings of the Joint Conference of the 15th Language Resources and Evaluation Conference, and, the 30th International Conference on Computational Linguistics, Torino, Italy. European Language Resources Association.

Abhijnan Nath, Sheikh Mannan, and Nikhil Krishnaswamy. 2023. Axomiyaberta: A phonologically-aware transformer model for assamese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11629–11646.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Gabriele Paul. 1993. Approaches to abductive reasoning: an overview. *Artificial intelligence review*, 7(2):109–152.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Singapore. Association for Computational Linguistics.

Rajat Raina, Andrew Y Ng, and Christopher D Manning. 2005. Robust textual inference via learning and abductive reasoning. In *AAAI*, pages 1099–1105.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1700–1716.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. Learning by distilling context.

Aleksandr Sokolov. 2012. *Inner speech and thought*. Springer Science & Business Media.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. *arXiv preprint arXiv:2206.11083*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jennifer Tracey, Ann Bies, Jeremy Getman, Kira Griffitt, and Stephanie Strassel. 2022. A study in contradiction: Data and annotation for aida focusing on informational conflict in russia-ukraine relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1831–1838.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 39–49, Nancy, France. Association for Computational Linguistics.

Marc Vilain, John D 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*.

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)*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284.

Kazeto Yamamoto, Naoya Inoue, Kentaro Inui, Yuki Arase, and Jun'ichi Tsujii. 2015. Boosting the efficiency of first-order abductive reasoning using pre-estimated relatedness between predicates. *International Journal of Machine Learning and Computing*, 5(2):114–120.

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

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland. Association for Computational Linguistics.

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

Jin Zhao, Nianwen Xue, and Bonan Min. 2023a. Cross-document event coreference resolution: Instruct humans or instruct GPT? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 561–574, Singapore. Association for Computational Linguistics.

Wenting Zhao, Justin T Chiu, Claire Cardie, and Alexander M Rush. 2023b. Abductive commonsense reasoning exploiting mutually exclusive explanations. *arXiv preprint arXiv:2305.14618*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Package and Pre-/postprocessing Details

We use the pretrained Longformer-base weights as they appear in the HuggingFace library.[4] For LLaMA 2-7B-Chat, we use the downloaded weights from Meta[5] which were then converted to the HuggingFace Transformers format for their pretrained libraries. For zero-shot evaluation using GPT-3.5-Turbo, we use OpenAI's `completions` API gateway.[6] We use the NLTK library[7] for tokenization when generating the rationale statistics at the token-level. For lemma-based heuristic candidate event generation, we use the popular spaCy Lemmatizer pipelines[8]. For getting the final coreference clusters after creating the affinity graph (post-transitive closure), we use the CoVal coreference scorer (Moosavi et al., 2019).

B Further Details on Motivation for Design Choices

Exclusion of repulsive regularization during ROEC Since our rationales are abductive in nature and have been conditioned on the gold coreference labels, it is likely that the step-by-step reasoning contains valid reasoning steps in support of the label, regardless of the actual status of the label (coreferent or not). Such informativeness of intermediate reasoning steps have also been observed in previous work (Wiegreffe et al., 2022) that suggests that a general purpose LLM can still generate plausible explanations given a task, even if it displays sub-par performance on the task itself especially in zero-shot evaluations.

Since our rationales incorporate multiple angles of reasoning about coreference, we hypothesize that the regularization process should only reward event representations that form similar clusters while events from separate clusters should remain unrewarded during training. This is because rationales for these events could still contain plausible hypotheses consistent with the cluster label. Therefore, we do not include a repulsive regularization component commonly used for creating separable clusters in pretraining event coreference models as seen in Kenyon-Dean et al. (2018) and Held et al. (2021).

Choice of temperate parameter The temperature we use for FTR generation (0.7) is the default value in LLaMA 2-7B-Chat. This creates controlled and focused responses to prompts without

[4] https://huggingface.co/allenai/longformer-base-4096

[5] https://ai.meta.com/llama/

[6] https://platform.openai.com/docs/guides/text-generation/completions-api

[7] https://www.nltk.org/api/nltk.tokenize.html

[8] https://spacy.io/api/lemmatizer

being fully deterministic. A conservative sampling strategy helps keep diversity of tokens in the rationales without added randomness that could negate the gold label or generate out-of-context outputs. Initial experimentation demonstrated 0.7 to be a reasonable value and a test of temperature values was determined to be out of scope for the paper.

Heuristic threshold optimization The threshold for the heuristic used to preprocess the data according to Ahmed et al. (2023)'s method is optimized using the validation data to minimize the loss of truly coreferent pairs, while pruning the large number of non-informative true negative mentions during candidate selection. A lower threshold also lets us select lexically misleading mention pairs that are actually not coreferent (hard negatives), that the heuristic fails on. These pairs are frequently seen in CDCR corpora, particularly in GVC (Vossen et al., 2018).

Since our rationales are defined as a one-to-one map with the corresponding mention pairs, such hard negatives paired with their rationales provide a richer signal during ROEC optimization. This helps the student model distinguish between pairs in the coreference cluster graph. Due to the sparsity of CDCR links (Bugert et al., 2021b), the heuristic maintains a class balance between positive and negative pairs without resorting to a document clustering step. This allows pairwise classifiers to learn more efficiently from a relatively balanced class distribution.

C Additional Results Tables

In Tables 5–7, we present commonly used CDCR metrics (Moosavi et al., 2019), comparing our systems' performances on ECB+, GVC, and AIDA Phase 1 to zero-shot evaluations, and previous baselines where available, to aid in future comparisons. We show MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$, and CoNLL F_1 (the average of MUC, B^3 and $CEAF_e$ F_1) scores. Although we do not always beat previous evaluations on EBC+ and GVC on all metrics, we frequently either do with at least one of our methods, or remain extremely competitive, to within at most 5 F1 points—the difference is usually within fractions of a point.

In coreference tasks, choice of metric reflects heavily in the results. For instance, almost 33% of the ECB+ dataset across all three splits consists of singleton mentions, and MUC score is not as sensitive to the presence of singletons as B^3. On the other hand, $CEAF_e$'s alignment algorithm tends to ignore correct coreference decisions when response entities are misaligned (Moosavi and Strube, 2016).

D Zero-Shot Prompt Design

The prompt formats used for zero-shot evaluation of LLaMA 2-7B-Chat and GPT-3.5-Turbo are given below.

LLaMA 2-7B-Chat Zero-shot Prompt Format

SYSTEM_PROMPT: Think step by step. You are a coreference annotator and you have to make a decision about two events marked with <m> and </m> tokens. You are given two sentences. Answer in one word if they are talking about different events or the same event.
USER_PROMPT: Sentence 1 is: {sentence_1}. Sentence 2 is: {sentence_2}. Your answer:

GPT-3.5-Turbo Zero-shot Prompt Format

SYSTEM_PROMPT: You are a coreference annotator and you have to make a decision about two events marked by <m> and </m> tokens. You are given two sentences. Answer in one word if they are talking about the same event: that is, if they are coreferent.
USER_PROMPT: sentence_1: {sentence_1} sentence_2: {sentence_2}

E Quality of Free Text Rationales

To evaluate the quality of information presented through the rationales, we presented human evaluators with a set of questions about a small sample of generated FTRs. The questions were inspired by Wiegreffe et al. (2021) and sought to assess factors like fact content, relevancy, plausibility, and the quality of the reasoning process demonstrated in the written output, according to humans.

Four evaluators (all adult English speakers) took a survey containing pairs of event mentions from two different documents (6 pairs each drawn from the ECB+ and GVC corpora), the ground truth label (which was also given to LLaMA 2-7B-Chat for generation), and the generated inner monologue

	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Caciularu et al. (2021)	**87.1**	89.2	**88.1**	84.9	87.9	86.4	83.3	81.2	82.2	85.6
Held et al. (2021)	87.0	88.1	87.5	85.6	87.7	86.6	80.3	**85.8**	82.9	85.7
Yu et al. (2022)	88.1	85.1	86.6	**86.1**	84.7	85.4	79.6	83.1	81.3	84.4
Ahmed et al. (2023) (w/o oracle)	80.0	87.3	83.5	79.6	85.4	82.4	83.1	75.5	79.1	81.7
LLaMA 2-7B-Chat (zero-shot)	84.2	76.3	80.1	82.7	73.2	77.7	67.5	77.2	72.0	76.6
GPT-3.5-Turbo (zero-shot)	81.7	81.0	81.4	81.0	78.6	79.8	76.1	77.0	76.5	79.2
$Long_{\text{paired}}$	81.5	84.1	82.8	81.1	82.4	81.8	79.4	76.5	77.9	80.8
$Long_{\text{+ROEC,-KD}}$ (ours)	79.4	**92.4**	85.4	79.8	**93.1**	85.9	**89.1**	76.1	82.1	84.5
$Long_{\text{-ROEC,+KD}}$ (ours)	78.2	90.6	83.9	79.4	90.2	84.4	87.9	75.4	81.2	83.2
$Long_{\text{+ROEC,+KD}}$ (ours)	84.1	92.0	87.9	82.4	91.7	**86.8**	88.9	80.5	**84.5**	**86.4**

Table 5: ECB+ test set evaluation results.

	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Bugert et al. (2021a)	78.1	66.3	71.7	73.6	49.9	59.5	38.2	60.9	47.0	59.4
Held et al. (2021)	91.8	91.2	91.5	82.2	**83.8**	83.0	75.5	**77.9**	**76.7**	**83.7**
Ahmed et al. (2023) (w/o oracle)	84.0	91.1	87.4	79.0	76.4	77.7	69.6	52.5	59.9	75.0
LLaMA 2-7B-Chat (zero-shot)	**93.9**	84.3	88.8	89.5	38.1	53.4	28.9	54.9	37.9	60.0
GPT-3.5-Turbo (zero-shot)	88.6	81.9	85.1	82.6	35.4	49.6	27.1	41.1	32.7	55.8
$Long_{\text{paired}}$	89.6	92.2	90.8	86.4	66.7	75.3	66.2	59.2	62.5	76.2
$Long_{\text{+ROEC,-KD}}$ (ours)	91.9	92.5	92.2	**86.8**	75.3	80.6	66.9	65.3	66.1	79.6
$Long_{\text{-ROEC,+KD}}$ (ours)	91.3	95.1	**93.2**	86.0	79.2	82.5	**76.6**	65.5	70.6	82.1
$Long_{\text{+ROEC,+KD}}$ (ours)	91.6	**94.2**	92.9	86.7	82.1	**84.3**	75.8	68.1	71.7	83.0

Table 6: GVC test set evaluation results.

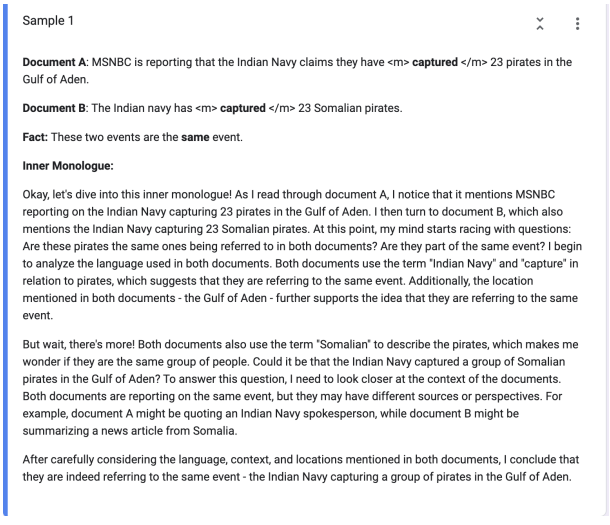
	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
LLaMA 2-7B-Chat (zero-shot)	**69.1**	65.9	67.5	**60.0**	39.7	47.8	45.7	**51.0**	48.2	54.5
GPT-3.5 Turbo (zero-shot)	55.6	71.4	62.5	54.4	57.7	56.0	63.0	42.7	50.9	56.5
$Long_{\text{paired}}$	57.5	76.4	65.6	54.4	62.4	58.1	68.0	44.4	53.7	59.1
$Long_{\text{+ROEC,-KD}}$ (ours)	56.9	87.7	69.1	52.3	73.7	61.2	83.1	47.6	60.5	63.6
$Long_{\text{-ROEC,+KD}}$ (ours)	60.2	88.1	71.5	53.0	73.3	61.5	81.7	48.8	61.1	64.7
$Long_{\text{+ROEC,+KD}}$ (ours)	60.6	**90.6**	**72.6**	53.3	**81.5**	**64.5**	**85.3**	50.0	**63.0**	**66.7**

Table 7: AIDA Phase 1 test set evaluation results.

FTR (see Fig. 4). They were asked to answer seven multiple choice questions for each sample, designed to explore various aspects of the inner monologue-based explanation.

The questions included:

- **Factuality**: How factual is this Inner Monologue-based explanation? (*generally true*/*partially true*/*generally false*/*not enough information*)
- **Relevance**: Is the Inner Monologue-based explanation relevant to the context? (*yes*/*no*)
- **New Information**: Does the Inner Monologue sample provide new facts, information, or reasoning not stated in the pair of documents explicitly? (*yes*/*no*)


Sample 1

Document A: MSNBC is reporting that the Indian Navy claims they have <m> **captured** </m> 23 pirates in the Gulf of Aden.

Document B: The Indian navy has <m> **captured** </m> 23 Somalian pirates.

Fact: These two events are the **same** event.

Inner Monologue:

Okay, let's dive into this inner monologue! As I read through document A, I notice that it mentions MSNBC reporting on the Indian Navy capturing 23 pirates in the Gulf of Aden. I then turn to document B, which also mentions the Indian Navy capturing 23 Somalian pirates. At this point, my mind starts racing with questions: Are these pirates the same ones being referred to in both documents? Are they part of the same event? I begin to analyze the language used in both documents. Both documents use the term "Indian Navy" and "capture" in relation to pirates, which suggests that they are referring to the same event. Additionally, the location mentioned in both documents - the Gulf of Aden - further supports the idea that they are referring to the same event.

But wait, there's more! Both documents also use the term "Somalian" to describe the pirates, which makes me wonder if they are the same group of people. Could it be that the Indian Navy captured a group of Somalian pirates in the Gulf of Aden? To answer this question, I need to look closer at the context of the documents. Both documents are reporting on the same event, but they may have different sources or perspectives. For example, document A might be quoting an Indian Navy spokesperson, while document B might be summarizing a news article from Somalia.

After carefully considering the language, context, and locations mentioned in both documents, I conclude that they are indeed referring to the same event - the Indian Navy capturing a group of pirates in the Gulf of Aden.

Figure 4: Rationale sample presented to evaluators.

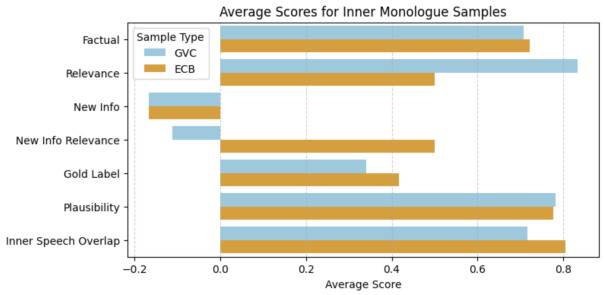


Figure 5: Average scores for inner monologue samples generated from ECB+ and GVC.

- **New Information Relevance**: If you answered yes to the above question, is the new information or reasoning relevant to justifying the facts about the events? (*yes/no/not enough information*)
- **Gold Label**: How much information does the Inner Monologue sample have to justify the facts about the two events? (*enough/not enough/more than enough/can't say*)
- **Plausibility**: Is the Inner Monologue sample acceptable or plausible considering the context? (*yes/no/can't say*)
- **Inner Speech Overlap**: If you were to use your own inner-monologue-based reasoning to arrive at the FACT, how much of an overlap does your thought-pattern have with the given Inner Monologue? (*high overlap/some overlap/minimal overlap/no overlap*)

Annotations were performed by members of the SIGNAL lab in the course of their normal duties. A different set of 4 annotators were used to assess samples from each corpus—2 male and 2 female in each set. Annotators had some casual exposure to the problem of CDCR but no other prior experience in the task. The survey was determined to be Not Human Subjects Research by the institutional review board.

The answers were then mapped to numerical values following the template of Wiegreffe et al. (2022). Yes/no answers were mapped to -1/1. Answers to the multiple-choice questions were mapped to -1 (negative valence), 0 (uninformative/neutral), 0.5 (partially positive valence, where relevant), or 1 (positive valence). Figure 5 shows average scores for inner monologue samples on the above questions.

FTRs generated from both datasets were rated as highly factual, showing that they were representing accurate information about the events concerned. GVC FTRs were rated as more highly relevant than ECB+ FTRs. This may reflect the low topic diversity of GVC (since all events concern gun violence, as long as the FTR remains on-topic, it remains relevant). Negative scores for new information on both datasets indicate a challenge in generating content not already mentioned in the source material. This suggests a potential area for improvement in terms of content generation fidelity. Where the FTR introduced new information, evaluators of ECB+ FTRs found this information more relevant than evaluators of GVC FTRs (this may also be a sparsity effect). FTRs from both datasets were rated as highly plausible, indicating similar levels of logical coherence in their inner monologue samples. When asked to assess the level of overlap between how the FTR proceeded and how they would think about the question if using inner speech, evaluators rated this highly.

We calculated Krippendorff's α (Krippendorff, 2011) as a metric of evaluator agreement. We found that $\alpha \approx .22$ for ECB+ FTRs and $\alpha \approx .06$ for GVC FTRs. While the average scores above indicate that the generated FTRs appear to contain information for coreference decisions that humans consider useful, the annotator agreement scores indicate the subjective nature of the evaluation task. Conditioning generation on the gold label, and providing the label to the annotators, places controls on human evaluation of rationales, since annotators tends to inject bias against rationales when they disagree with the gold label (Wiegreffe et al., 2022). Untrained humans often weigh the same information significantly differently in the same task (Zhao et al., 2023a).

F Okay, Let's Do This?

The title of our paper comes from the fact that LLaMA 2-7B-Chat begins its FTRs with a "chatty" introductory sentence before starting to generate content regarding the event pair. "Okay, let's do this!" is one of the most frequent (and amusing) introductory sentences occurring in our FTR corpus.