Seeing the Same Story Differently: Framing-Divergent Event Coreference for Computational Framing Analysis

Anonymous ACL submission

Abstract

News articles often describe the same real-world event in strikingly different ways, shaping perception through framing rather than factual disagreement. However, traditional computational framing approaches often rely on coarse-grained topic classification, limiting their ability to capture subtle, event-level differences in how the same occurrences are presented across sources. We introduce **Fr**aming-divergent **E**vent **Co**reference (FRECO), a novel task that identifies pairs of event mentions referring to the same underlying occurrence but differing in framing across documents to provide a event-centric lens for computational framing analysis. To support this task, we construct the high-agreement and diverse FRECO corpus. We evaluate the FRECO task on the corpus through supervised and preference-based tuning of large language models, providing strong baseline performance. To scale beyond the annotated data, we develop a bootstrapped mining pipeline that iteratively expands the training set with high-confidence FRECO pairs. Our approach enables scalable, interpretable analysis of how media frame the same events differently, offering a new lens for contrastive framing analysis at the event level. The dataset and code will be made publicly available. [1]

1 Introduction

Media framing is the strategic act of emphasizing certain aspects of an issue while downplaying others, often to promote a particular narrative or interpretation (Entman, 1993). Consider the following two sentences:

(1) The officer acted decisively to ***neutralize the threat*** ... The officer ***opened fire*** on the unarmed man...

Both describe the same real-world police shooting event, but presenting two radically different stories. The first frames the shooting event from a security-focused perspective, emphasizing the necessity of the officer's action. The second adopts a justice-focused lens, highlighting the harm and moral implications. Both descriptions are factually grounded, yet they diverge in lexical framing, emotional valence, and moral attribution. These framing differences profoundly shape how audiences interpret events.

As digital media platforms continue to multiply and dominate public discourse, the demand for automated, reliable methods of framing analysis has become increasingly urgent. A central challenge is capturing how same real-world event can evolve into sharply divergent narratives without changing its factual basis. Existing computational framing methods often rely on topic modeling or predefined frame taxonomies, which impose a limited set of labels onto the data. These frame inventories are typically domain- and culture-specific, making them difficult to generalize across issues. For example, frames used to analyze immigration coverage may not meaningfully apply to reporting on gun violence.

In this paper, we introduce **Fr**aming-divergent **E**vent **Co**reference (FRECO), a novel task that identifies pairs of event mentions referring to the same underlying occurrence with contrastive framing. FRECO captures lexical, causal, and perspective contrasts between coreferential event mention pairs, enabling more nuanced and interpretable framing analysis at the event level. FRECO ties event coreference to narrative contrast, offering a new scalable computational lens on how media shape reality through framing. Unlike traditional coreference tasks, FRECO treats variation in argument structure, granularity, and perspective not as noise, but as the signal. In contrast to current framing research, our event-centric approach operates at the level of real-world events, which remain interpretable and stable across domains. By anchoring framing analysis in event structure, our

[1] Anonymized for review.

method offers a more scalable and systematic way to study framing variation across topics, genres, and cultural contexts.

Our goal is not only to classify whether a pair qualifies as a FRECO instance, but to develop models that can surface such framing-divergent coreferential pairs at scale in real world reporting. This framing-aware coreference detection opens new possibilities for contrastive framing analysis across large corpora, making FRECO both a theoretically rich and practically impactful task.

To support this task, we construct the FRECO Corpus, a dataset of framing-divergent coreferential event pairs drawn from ideologically diverse news coverage of contentious events. We leverage Large Language Models (LLMs) to systematically extract FRECO pairs. We evaluate multiple modeling strategies, and find that combining Supervised Fine-Tuning (SFT) with Direct Preference Optimization (DPO) yields the best overall performance. Incorporating structured event representations further enhances model accuracy.

To scale beyond the gold annotated data, we introduce a bootstrapping mining pipeline. Starting from a filtered pool of semantically similar event pairs, we iteratively apply our trained classifier to identify high-confidence FRECO instances. Each round expands the training set with newly mined pairs, enabling our model to uncover framing-divergent coreference at scale. This pipeline allows us to transition from classification to retrieval, offering a scalable path for contrastive framing analysis in the wild.

Our contributions are as follows:

1. We define the novel task of FRECO and construct the first high-quality FRECO corpus grounded in ideologically diverse news coverage.
2. We develop the first NLP system for FRECO by finetuning LLMs with SFT and DPO, and demonstrate their complementary effectiveness in capturing framing divergence.
3. We introduce a bootstrapping mining pipeline to scale FRECO identification beyond gold annotations, enabling high-confidence extraction of framing-divergent event pairs at scale.

2 The FRECO Task

Definition We define FRECO as the task of identifying pairs of event mentions that refer to the same real-world occurrence with contrastive framing. These divergences may arise through lexical choice, causal attribution, emotional valence, specificity, or narrative perspective (Goffman, 1974).

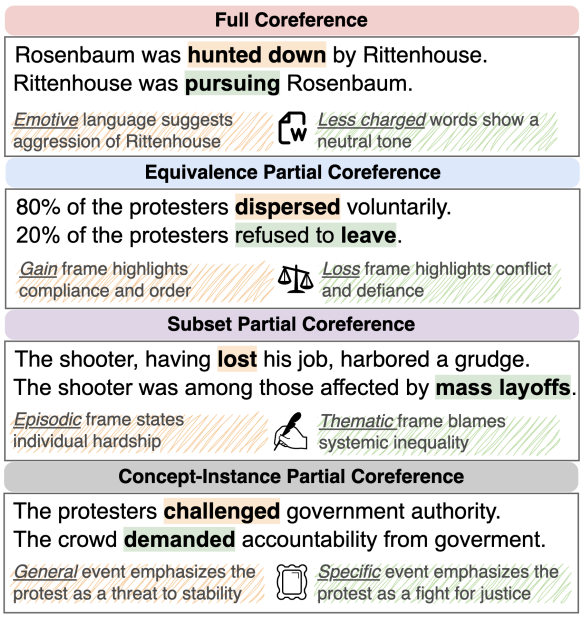

Figure 1: Examples of FRECO Pairs. Event trigger words are highlighted.

FRECO builds on the relaxed identity notion of event hoppers (Mitamura et al., 2017) in Cross-Document Event Coreference (CDEC) research and includes both fully and partially coreferential event mentions (Hovy et al., 2013). As illustrated in Figure 1, the FRECO dataset captures a spectrum of coreference types, from full coreference with tonal contrast (*hunted down* vs. *pursued*), to logical equivalence with divergent framing (*dispersed* vs. *refused to leave*), to subset or subevent relations (*lost his job* vs. *mass layoffs*), and abstraction-level differences (*challenged authority* vs. *demanded accountability*). Despite differing structure or emphasis, all of these event pairs refer to the same underlying real-world event but highlight different aspects of it, revealing how narratives diverge.

FRECO as Classification Task FRECO is framed as a classification problem. Let e_1 and e_2 be two event mentions drawn from one or more documents on the same topic. The FRECO task is to predict a binary label $y \in \{0, 1\}$ where `y = 1` if e_1 and e_2 refer to the same real-world occurrence but contrast in framing, and `y = 0` otherwise.

FRECO as Mining Task We also operationalize FRECO as a retrieval task to support large-

scale mining. Given a large set of event mentions $\mathcal{E} = \{e_1, e_2, ..., e_n\}$, our goal is to retrieve a subset $\mathcal{P} \subset \mathcal{E} \times \mathcal{E}$ such that each $(e_i, e_j) \in \mathcal{P}$ satisfies the FRECO condition. To scale this, we approximate retrieval using a hybrid pipeline: we first use cross-encoder embedding similarity to generate a filtered candidate pool, and then apply our fine-tuned FRECO classifier $f_\theta(e_i, e_j) \rightarrow [0, 1]$ to assign soft scores. We use these scores to iteratively mine high-confidence FrECo instances, starting with a small annotated seed set and expanding the dataset over multiple rounds. In each round, newly mined examples are added to the training data, allowing the model to improve its ability to surface framing-divergent coreference in the wild. This bootstrapped approach enables FRECO to transition from a classification task to a scalable retrieval pipeline, supporting contrastive framing analysis across large and diverse document collections.

3 The FRECO Corpus

We construct our dataset on top of the Richer Event-CorefBank (RECB) (Zhao et al., 2025), a cross-document event coreference dataset spanning four contentious topics: the 2024 Putin's election win (PUTIN), the Al-Shifa hospital raid (AL-SHIFA), the 2019 July 1 Hong Kong protest (HONGKONG), and the Kyle Rittenhouse shooting (RITTENHOUSE). Sourced from ideologically and geographically diverse media outlets, RECB is well-suited for framing analysis, capturing contrasting narratives across polarized perspectives. Rather than using RECB's original coreference annotations, we generate all possible event pairs within each topic with all event mentions from the RECB articles. These form the candidate pool for identifying framing-divergent coreferential event pairs.

3.1 Candidate Pair Selection

To efficiently identify promising FRECO candidates, we leverage the CDEC model (Yu et al., 2022), a state-of-the-art pairwise cross-encoder trained to detect cross-document event coreference. Although optimized for identity matching, CDEC similarity scores serve as a powerful proxy for surfacing event pairs that are semantically aligned but potentially divergent in framing.

We rank all event pairs by their cross-encoder similarity score and select the top-scoring candidates for human annotation. This approach prioritizes pairs that likely refer to the same underlying occurrence, increasing the density of valid FRECO examples in our dataset. For example, the mentions *self-defense* and *fired weapon* both refer to the same action in RITTENHOUSE, but differ in framing: the former conveys justification, while the latter remains neutral. Such examples reflect level of abstraction, granular and perspectival divergence rather than strict identity, and are highly ranked by CDEC model due to their semantic proximity.

3.2 Annotation

We hire two students from the computational linguistics program of a U.S.-based university for annotating the FRECO. Each annotator undergoes training on the definitions of contrastive framing and FRECO, ensuring they adhere to annotation guidelines (Appendix A.4).

Annotators are presented with candidate event mention pairs with their containing sentence contexts, ranked from highest to lowest in similarity by the CDEC scorer. They annotate each pair as FRECO or not. If a pair is marked as FRECO, annotators further determine whether each individual event's attitude is supportive or skeptical towards the main event of the article. (This additional annotation is used in downstream task evaluation in Appendix A.5).

To ensure consistency, annotators collaboratively review 100 training pairs, which include edge cases to clarify ambiguous instances and refine their understanding of the guidelines. They then double-annotate 200 pairs per subtopic, followed by joint adjudication of discrepancies. We measure Inter-Annotator Agreement using Cohen's κ, obtaining 0.76 for identifying FRECO and 0.81 for labeling individual event attitudes. It indicates high quality of our dataset and consistent application of guidelines.

3.3 Corpus Statistics

The FRECO corpus contains a total of 3,800 annotated event mention pairs spanning four contentious news topics: PUTIN (739 pairs), AL-SHIFA (1,356 pairs), HONGKONG (653 pairs), and RITTENHOUSE (1,052 pairs). Among these, 1,765 pairs are labeled as positive FRECO instances, constituting 46.5% of the total dataset. The remaining pairs are non-coreferential or lack framing divergence. This balanced distribution supports both classification and retrieval-based modeling for framing-divergent event coreference.

4 FRECO Pairwise Classification

We start by fine-tuning several language model–based classifiers on the FRECO corpus to classify event mention pairs that are coreferential yet differ in framing.

4.1 Data Preparation

Input Variants We evaluate the FRECO task using our annotated FRECO corpus and compare model performance across two dataset variants. The first variant contains sentences with tagged event mentions presented in their original document context. The second augments each example by explicitly highlighting event components using agents, patients, locations, and temporal arguments extracted by Semantic Role Labeling (SRL). For this version, models are prompted to compare the extracted components directly. To obtain these structured arguments, we treat each event trigger as a predicate and apply a transformer-based SRL parser[2] to identify its semantic roles.

Leave-One-Topic-Out Evaluation To evaluate cross-topic generalization, we adopt a variant of leave-one-group-out cross-validation, where each group corresponds to a topic. In each of the four folds, we hold out one topic as the test set and train on the remaining three topics, reserving 20% of the training data from each topic for development, enabling reliable early stopping and hyperparameter tuning without contaminating the evaluation topic. This setup ensures that models are evaluated on truly unseen events and framing strategies, rather than benefiting from topical or lexical memorization. Given the goal of mining framing-divergent event coreference across documents, this evaluation protocol tests a model's ability to generalize beyond surface-level similarity and capture framing-sensitive event representations across diverse geopolitical and ideological contexts.

4.2 Model Setup

As a baseline, we use recent open-source LLMs, specifically instruction-tuned Llama models (Dubey et al., 2024) at two different scales: `Llama-3.2-3B` and `Llama-3.1-8B`. We evaluate both dataset variants with these models. To establish the LLM baseline, we prompt each model directly for inference without additional tuning. We then explore a range of fine-tuning strategies on these models, including SFT, DPO, and their sequential combinations (SFT→DPO, DPO→SFT). Training details and prompts are provided in Appendix A.2.

As an additional baseline, we adopt the pairwise CDEC classifier architecture proposed by Yu et al. (2022) and adapt it to the FRECO classification. This model uses a $\text{RoBERTa}_{\text{BASE}}$ cross-encoder that jointly encodes the concatenated sentence pair with marked event mentions. The representations of the trigger tokens are aggregated into a unified feature vector, which is then used for binary classification. We follow the original implementation and settings described in Yu et al. (2022), and use the first version dataset variant for this experiment.

We additionally evaluate zero-shot performance using GPT-4, prompted on the first version of the dataset to assess its out-of-the-box ability to identify FRECO pairs without task-specific fine-tuning.

4.3 Results

Table 1 reports F1 scores on the FRECO classification task across four held-out test topics, comparing model performance under various finetuning strategies and model configurations. All models show a significant improvement after finetuning compared to the zero-shot baselines, indicating that finetuning is crucial for improving FRECO performance.

We observe that both DPO→SFT and SFT→DPO consistently achieve the best performance across most settings, outperforming standalone SFT or DPO models. This highlights the complementary strengths of supervised fine-tuning and preference-based optimization. Notably, DPO alone performs better than SFT alone in most cases, suggesting that pairwise, margin-based learning is particularly effective for this task in topics like PUTIN. This is likely due to the prevalence of hard negative examples in the dataset. Many non-FRECO pairs were ranked highly by the CDEC pairwise scorer, meaning they are semantically similar but ambiguous in terms of both coreference and framing divergence. Such examples benefit from the ranking signal provided by pair-wise, margin-based learning of DPO.

Larger models consistently outperform their smaller counterparts after finetuning, particularly when SRL features are used and advanced fine-tuning strategies are applied. Incorporating SRL-based structured event representations further improves performance in nearly all cases. This sug-

[2] https://huggingface.co/cu-kairos/propbank_srl_seq2seq_t5_large

Test Topic	Model	Inference(0-shot)	SFT	DPO	DPO→SFT	SFT→DPO
PUTIN	Llama-3.2-3B	43.31(±0.00)	75.21(±1.42)	77.81(±1.18)	**77.87(±2.05)**	77.54(±1.84)
PUTIN	Llama-3.1-8B	29.76(±0.00)	76.73(±1.20)	**79.51(±1.30)**	78.92(±0.77)	79.19(±0.63)
PUTIN	Llama-3.2-3B + SRL	46.48(±0.00)	76.59(±1.36)	**79.62(±1.04)**	79.37(±0.66)	78.85(±0.71)
PUTIN	Llama-3.1-8B + SRL	31.04(±0.00)	78.05(±1.59)	79.94(±0.89)	80.18(±0.81)	**80.55(±0.58)**
AL-SHIFA	Llama-3.2-3B	50.44(±0.00)	79.08(±2.87)	78.37(±1.14)	**79.92(±0.93)**	78.01(±0.65)
AL-SHIFA	Llama-3.1-8B	39.28(±0.00)	74.55(±1.54)	79.12(±1.76)	79.48(±0.80)	**79.64(±0.52)**
AL-SHIFA	Llama-3.2-3B + SRL	57.63(±0.00)	76.46(±1.22)	80.41(±1.10)	**80.56(±0.71)**	80.22(±0.77)
AL-SHIFA	Llama-3.1-8B + SRL	44.97(±0.00)	79.19(±1.32)	81.32(±1.29)	80.03(±1.90)	**81.38(±1.49)**
HONGKONG	Llama-3.2-3B	43.12(±0.00)	73.04(±1.35)	75.88(±1.44)	80.66(±0.92)	**80.79(±0.61)**
HONGKONG	Llama-3.1-8B	15.37(±0.00)	77.01(±2.41)	76.35(±1.52)	81.24(±0.88)	**81.47(±0.55)**
HONGKONG	Llama-3.2-3B + SRL	45.59(±0.00)	74.22(±1.26)	77.11(±1.17)	**82.02(±0.79)**	81.81(±1.68)
HONGKONG	Llama-3.1-8B + SRL	28.08(±0.00)	78.44(±1.68)	77.73(±1.23)	82.19(±1.83)	**82.36(±0.57)**
RITTENHOUSE	Llama-3.2-3B	59.23(±0.00)	74.11(±1.90)	77.43(±1.27)	82.46(±0.85)	**82.57(±0.73)**
RITTENHOUSE	Llama-3.1-8B	35.72(±0.00)	75.34(±1.66)	78.08(±1.41)	83.92(±1.69)	**84.07(±2.60)**
RITTENHOUSE	Llama-3.2-3B + SRL	61.88(±0.00)	79.56(±1.53)	78.94(±1.10)	**84.36(±0.72)**	84.11(±1.66)
RITTENHOUSE	Llama-3.1-8B + SRL	38.27(±0.00)	79.48(±1.74)	79.26(±1.24)	**84.95(±0.77)**	84.79(±0.55)

Table 1: Evaluation results on FRECO classification task across four test topics. We compare inference baselines and models trained under different strategies. F1 score (Mean ± Std) is reported.

	PUTIN	AL-SHIFA	HONGKONG	RITTENHOUSE
$\text{RoBERTa}_{\text{BASE}}$	78.14(±0.63)	78.86(±0.00)	80.71(±0.01)	78.10(±0.03)
GPT-4	51.57(±0.00)	62.53(±0.00)	57.56(±0.00)	64.31(±0.00)
Llama	80.55(±0.58)	81.38(±1.49)	82.36(±0.57)	84.95(±0.77)

Table 2: Result comparison of finetuned $\text{RoBERTa}_{\text{BASE}}$, GPT-4 and the best-performing Llama model configurations in Table 1.

gests that SRL contributes to better event understanding, making it beneficial for this task.

Table 2 shows that best fine-tuned LLMs consistently outperforms fine-tuned $\text{RoBERTa}_{\text{BASE}}$ cross-encoder classifier on all topics. GPT-4 underperforms both baselines, suggesting that zero-shot prompting is insufficient for capturing FRECO without task-specific adaptation.

These results demonstrate that FRECO classification benefits from both preference-based optimization and adding structured event components. The consistent gains across diverse test topics highlight the generalizability of our approach to FRECO across domains.

4.4 Error Analysis

We analyze the false negative and false positive pairs produced by the models that yields the best performance on our task. For false negative errors, the majority occur because the model struggles to determine whether two events are truly coreferential. A smaller portion results from the model failing to detect subtle framing differences. For instance, when framing valence is not embedded directly in the event itself but instead emerges through its causal connection to another event, the framing shift becomes more nuanced. The model, relying primarily on surface-level event features, fails to capture implicit causal attributions, leading to missed FRECO pairs.

For false positive errors, the majority stem from the model mistakenly considering unrelated events as equivalent. This often happens when the events share a similar nature (e.g., two separate *deaths* within the same conflict) or when both sentences have strong opposing framing, leading the model to overgeneralize their connection. See the following examples for a more detailed analysis.

(2) Sentence A: Jurors listened to two weeks of dueling **portrayals** of Rittenhouse, with the prosecution depicting him as an aggressor and the defense portraying him as acting in self-defense.

Sentence B: In his opening argument on Tuesday during the trial of Kyle Rittenhouse for the shootings in Kenosha, Kenosha County Assistant District Attorney Thomas Binger **painted** a wild version of events that isn't even close to what's in the state's original criminal complaint against Rittenhouse.

Ground truth: **Postive**;
Finetuned Llama-3.1-8B Prediction: **Negative**;

Both events in (2) describe competing narratives in Rittenhouse's trial. *Dueling portrayals* in sentence A presents a balanced exchange between the prosecution and defense, implying neutrality. *Painted* in sentence B focuses solely on the bias of prosecution. Therefore this is a positive example of equivalent event with contrastive framing.

Finetuned `Llama-3.1-8B` failed likely due to its inability to recognize the same underlying event. Sentence A includes both the prosecution and defense as active participants, while Sentence B highlights only the prosecution. The model treated them as distinct events.

(3) Sentence A: Fighting raged on Saturday around Gaza's main hospital where Israel says it has so far killed more than 170 gunmen in an extensive raid , which the Palestinian health ministry says has also resulted in **death_EVENT** of a patient .

Sentence B: Yuval Nir, a resident of Kfar Etzion and a dedicated soldier, **fell_EVENT** in battle in Gaza during military operations in Gaza.

Ground truth: **Negative**;
Finetuned `Llama-3.1-8B` Prediction: **Positive**;
Finetuned `Llama-3.1-8B` + SRL Prediction: **Negative**

Example (3) shows an false positive error, where the framing contrast is strong, but the underlying events are not actually coreferential: one refers to the death of a patient in Al-Shifa Hospital, while the other describes the death of a soldier in the battle field. There is no overlap in participants and context, making them entirely separate events.

Finetuned `Llama-3.1-8B` failed likely due to the framing contrast between the two articles is highly detectable. Sentence A presents a skeptical stance toward Israel's raid on Al-Shifa Hospital, highlighting civilian casualties, while Sentence B portrays the soldier's death as honorable, aligning with a supportive perspective for the military operation. While the framing differences are clear, the model appears to have overgeneralized based on narrative contrast rather than event equivalence, misidentifying framing opposition as a sufficient condition for event similarity. The SRL-enhanced model did not make this mistake, as it explicitly compared the experiencer of the event — identifying that the patient in Sentence A and the soldier in Sentence B are different entities. By focusing on argument structures rather than just surface-level framing, the SRL model correctly determined that these two events are unrelated.

This suggests that our model may be overweighting discourse-level framing cues while underweighting equivalence and participant alignment, leading to false positives in cases where differently framed deaths are not actually referring to the same event. Addressing this requires refining the model's argument structure alignment.

5 Bootstrapped FRECO Mining

To scale beyond the annotated FRECO corpus, we design a bootstrapped mining framework that leverages a small set of gold-labeled event pairs to iteratively expand high-confidence FRECO instances from the full RECB article collection.

Candidate Generation We begin with annotated FRECO pairs, of which 80% are used for training and 20% are held out as a development set for validation. The full RECB corpus yields approximately 4.87 million candidate pairs when naively pairing all events within each topic. To reduce this space, we use CDEC pairwise scorers (Yu, 2023) to rank all possible event pairs by similarity for each topic. We discard all easy negative pairs with similarity scores below 0.3 (elbow point in the similarity distribution), resulting in around 45k candidate pairs. This pool includes the original training data but excludes both the development set and the long tail of low-similarity examples.

Bootstrapping Procedure Using the four best FRECO classifiers from Section 4, we score the candidate pairs and retain those with a model confidence above 0.9 as high-confidence FRECO pairs. This results in 4,213 pseudo-labeled seed pairs for bootstrapping. In each subsequent round, we expand the training set by combining the gold examples in training set with the newly mined pseudo-labeled pairs and retrain the model (yielding FRECO classifiers in round 1, 2 and so on). After retraining, the updated model is applied again to the candidates pool to re-score candidate pairs. We gradually lower the threshold for inclusion in each round to expand the set of mined examples, as shown in Table 3.

Stopping Criteria We terminate bootstrapping after Round 3 based on multiple convergence indicators. First, the number of newly mined positive pairs drops sharply after Round 3. Second, the validation loss plateaus between Rounds 3 and 4 and begins to increase in Round 5, indicating potential overfitting or the introduction of noisy pairs. Third, the Jaccard similarity between newly mined sets in successive rounds steadily decreases, signaling that the model is exploring increasingly dissimilar and potentially less reliable regions of the candidate space. Fourth, we observe a degradation in precision based on manual review of randomly sampled pairs near the model's threshold that ambiguous or noisy pairs begin to dominate.

Round	Threshold	+ Pairs	+ Pos Pairs	Cumul.	Cumul. Pos	Jaccard	Val. Loss
Seed (Gold only)	–	–	–	3,040	1,765	–	0.410
Bootstrapping Init	0.90	4,213	1,127	7,253	2,892	–	0.382
Round 1	0.85	8,632	3,287	15,885	6,179	0.58	0.340
Round 2	0.83	4,954	1,683	20,839	7,862	0.30	0.332
Round 3	0.82	2,210	596	23,049	8,458	0.19	0.331
Round 4	0.81	1,115	223	24,164	8,681	0.12	0.328
Round 5	0.80	2,030	263	26,194	8,944	0.08	0.337

Table 3: Bootstrapped mining results across iterations. Each round lowers the model prediction threshold and adds newly mined high-confidence FRECO pairs to the training set. **Threshold** refers to the confidence score cutoff for selecting positive pairs. **+ Pairs** indicates the total number of newly mined pairs added in that round, while **+ Pos Pairs** specifies how many of them were labeled as positive FRECO pairs. **Cumul.** reports the cumulative training set size, including the original 3,040 gold-labeled examples. **Jaccard** measures the similarity between newly mined sets in consecutive rounds. **Val. Loss** is the average cross-entropy loss on a held-out validation set of 760 pairs.

Result As shown in Table 3, by the end of Round 3, the bootstrapping process added 6,693 new positive FRECO pairs, augmenting the original 1,765 gold positive pairs and substantially expanding the pool of high-confidence framing-divergent examples. The final mined dataset achieves 88% recall with respect to the original gold-labeled FRECO pairs. Precision is estimated at 70.5% based on human evaluation of 200 randomly sampled mined examples. Since true recall over the full corpus is unknowable, we treat this as an upper-bound estimate of mining quality. These results demonstrate that our semi-automatic bootstrapped framework effectively identifies new positive FRECO pairs at scale by leveraging a small gold set and iterative expansion.

6 Related Work

6.1 Event Coreference Definitions

Event coreference aims to determine whether two or more mentions refer to the same real-world occurrence, typically based on alignment of attributes like trigger, participants, time, and location (Hovy et al., 2013). The strictest definition, full event coreference, links only mentions that match across all dimensions, as in ACE (Linguistic Data Consortium, 2005), OntoNotes (Pradhan et al., 2007), and EventCorefBank (ECB+) (Cybulska and Vossen, 2014).

More flexible definitions, such as partial or quasi coreference (Hovy et al., 2013; Araki et al., 2014), capture hierarchical or gradable relationships such as subevent (e.g., a bombing as part of a larger terrorist attack) or membership (e.g., one protest among many) or concept-instance (e.g., arresting protesters as a concrete instance of abstract event crackdown on dissent), where events share substantial semantic overlap without being strictly identical. TAC KBP's event hoppers (Mitamura et al., 2017) further relax constraints, clustering mentions that are intuitively related despite differences in arguments, granularity, or realis status. Our work adopts this more relaxed view, reflecting the discourse-driven and framing-sensitive nature of event interpretation.

6.2 Computational Framing Methods

Computational Framing analysis has seen considerable attention within NLP, often focusing on identifying topic-level frames in news coverage. A variety of datasets approximate framing via topic, such as the Media Frames Corpus (Card et al., 2015) and its extensions (Piskorski et al., 2023; Ajjour et al., 2019; Mendelsohn et al., 2021), BU-NEmo (Reardon et al., 2022), as well as issue-specific corpora including the Gun Violence Frame Corpus (Liu et al., 2019), VoynaSlov (Park et al., 2022), and the stereoimmigrants dataset (Sánchez-Junquera et al., 2021). These resources facilitate various computational methods, spanning topic modeling (DiMaggio et al., 2013; Nguyen et al., 2015), unsupervised learning (Burscher et al., 2016), semantic parsing (Ziems and Yang, 2021), and finetuned smaller language models (Mendelsohn et al., 2021).

In recent years, a more event-centric lens on framing emphasizes how events and their relations—temporal, causal, or otherwise—shape public interpretation of issues. For instance, Liu et al. (2023) aligns news articles covering the same story to highlight the selection of partisan events. Das et al. (2024) clusters event relations into narratives to reveal frames. Zhao et al. (2024) compares context events of the main event to reveal the selection/omission aspect of framing. While these ap-

proaches have substantially contributed to studying framing automatically, They often overlook subtle differences in how the same real-world events can be framed positively or negatively without changing the underlying facts.

Our work bridges these gaps by focusing on coreferential events presented with varying perspectives and connotations. We construct a new dataset of such events, expanding beyond traditional text-span-based framing corpora and predefined topical categories. Additionally, we explore finetuning LLMs with different strategies to assess their effectiveness in capturing contrastive framing distinctions.

6.3 Framing Conceptualization in NLP

Emphasis and word choice framing have been the primary focus of most NLP research on media framing. However, existing work often simplifies the nuanced theoretical concept of framing. Many studies model emphasis framing using a limited set of predefined topics as proxies for frames (Sarmiento et al., 2022; Nicholls and Culpepper, 2020), while more recent work focuses on event selection and omission (Liu et al., 2023; Zhao et al., 2024). Yet, these approaches typically overlook subtle variations in valence, participant portrayal, and narrative focus within descriptions of the same real-world event. Similarly, studies on word choice framing have explored metaphorical language (Mendelsohn et al., 2020; Card et al., 2022), modifiers (Kwak et al., 2020; Jing and Ahn, 2021), and evaluative adjectives (Luo et al., 2024), but often operate in isolation without a unified theoretical framework, limiting generalizability. Labels are frequently inferred using heuristics such as collocations or semantic similarity (Sheshadri et al., 2021).

Our work systematically conceptualize emphasize and word choice framing within an event-based framework, as shown in first and forth example pair in Figure 1, allowing analysis of how language choices—such as action verbs, participant descriptions, temporal markers, and locations—shape perception and emotional response. This approach supports a more integrated and theoretically grounded analysis of framing in media narratives.

Equivalence and narrative framing remain underexplored in computational framing research. Existing work on equivalence framing often relies on corpus-level statistics (Luo and Huang, 2022; Chen et al., 2022), FrameNet-based frame comparisons (Postma et al., 2020), or handcrafted lexicons for domain-specific tasks like phishing detection (Dalton et al., 2020). However, these approaches are limited in scope and lack generalizable frameworks. Our FRECO system captures equivalence framing by identifying coreferential event pairs that differ in gain/loss framing, as illustrated in the second example in Figure 1.

Similarly, prior work on narrative framing has focused on identifying characters, motives, and plot structures (Mendelsohn et al., 2021; Pan et al., 2023), rarely connects these elements to established frame schemas such as the episodic–thematic distinction (Otmakhova et al., 2024). FRECO supports this narrative dimension by automatically extracting and contrasting narrow episodic and broad thematic event descriptions, aligning with Iyengar (1993)'s distinction between episodic and thematic frames. This is exemplified in the third example pair in Figure 1.

7 Conclusion

We introduce an event-centric approach to contrastive framing analysis that moves beyond topic-based models and rigid frame inventories. Our framework centers on FRECO, a new task that captures how media report the same underlying events with divergent framing through shifts in lexical choice, emotional valence, causal attribution, narrative perspective, and emphasis. To support this task, we construct the FRECO corpus, featuring annotated event pairs from ideologically diverse coverage of contentious topics. We fine-tune LLMs on this corpus to build effective FRECO classifiers and scale the task via a bootstrapped mining pipeline. This iterative process extracts high-confidence FRECO pairs from millions of candidates, enabling large-scale framing analysis with high precision and cross-domain generalizability.

This work contributes a scalable and interpretable methodology for media transparency, offering practical applications for news aggregators, bias-detection systems, and computational tools designed to counteract manipulation and polarization. Beyond NLP, our framework opens new possibilities for journalism and communication scholars by enabling large-scale, event-grounded exploration of framing strategies across geopolitical and cultural contexts.

Limitations

Topical and Media Coverage The FRECO corpus currently spans only four contentious topics, which may not represent the full range of framing techniques. Future work can broaden its topical scope to capture a wider spectrum of framing strategies. Including less polarized or non-political domains, such as health, climate, or technology reporting, could uncover more subtle or culturally contingent framing differences. Additionally, expanding to multilingual corpora or alternative media types (e.g., social media, podcasts, or international news outlets) would enhance the robustness and cross-cultural applicability of framing-divergent coreference analysis. This would also support more inclusive and globally relevant framing studies.

Toward Interpretable Framing Categorization Our current approach detects whether two event mentions differ in framing, but it does not yet characterize how they differ. Future work could enhance interpretability by categorizing framing contrasts along established dimensions, such as tone, moral attribution, causal focus, or frame theme. This could be achieved through framing-aware causality research, lexicon-based analysis, or supervised labeling of frequent framing patterns in mined pairs. Such an extension would deepen the connection between computational output and social science framing theory, enabling richer analyses of how divergent narratives emerge around the same events.

Ethical Considerations

The ability to detect contrastive framing in news articles has broad implications, including media bias analysis, misinformation detection, and propaganda studies. While our intent is to support academic and journalistic transparency, we recognize that such methods could be misused to selectively discredit certain narratives or manipulate public perception. To prevent misuse, we advocate for responsible use of framing analysis.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Conference on Empirical Methods in Natural Language Processing*.

Jun Araki, Eduard Hovy, and Teruko Mitamura. 2014. Evaluation for partial event coreference. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 68–76, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bjorn Burscher, Rens Vliegenthart, and Claes H. de Vreese. 2016. Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5):530–545.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119.

Jieyu Chen, Kathleen Ahrens, and Chu-Ren Huang. 2022. Framing legitimacy in CSR: A corpus of Chinese and American petroleum company CSR reports and preliminary analysis. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 24–34, Marseille, France. European Language Resources Association.

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).

Adam Dalton, Ehsan Aghaei, Ehab Al-Shaer, Archna Bhatia, Esteban Castillo, Zhuo Cheng, Sreekar Dhaduvai, Qi Duan, Bryanna Hebenstreit, Md. Mazharul Islam, Younes Karimi, Amirreza Masoumzadeh, Brodie Mather, Sashank Santhanam, Samira Shaikh, Alan Zemel, Tomek Strzalkowski, and B. Dorr. 2020. Active defense against social engineering: The case for human language technology. In *Symposium on the Theory of Computing*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth. Note: We use this library for dpo support.

Rohan Das, Aditya Chandra, I-Ta Lee, and Maria Leonor Pacheco. 2024. Media framing through the lens of event-centric narratives. Note: media framing through event perspective: Rohan used event temporal relations.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606. Topic Models and the Cultural Sciences.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Robert Entman. 1993. Framing: Toward clarification of a fractured paradigm. *The Journal of Communication*, 43:51–58.

Erving Goffman. 1974. Frame analysis: An essay on the organization of experience. *Philosophy and Phenomenological Research*, 39(4):601–602.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia. Association for Computational Linguistics.

Shanto Iyengar. 1993. Is anyone responsible?: How television frames political issues. *American Journal of Sociology*, 98(6):1459–1462.

Elise Jing and Yong-Yeol Ahn. 2021. Characterizing partisan political narrative frameworks about covid-19 on twitter. *Epj Data Science*, 10.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017.

Linguistic Data Consortium. 2005. Ace (automatic content extraction) english annotation guidelines for events. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, Kaijian Zou, Ruihong Huang, Nick Beauchamp, and Lu Wang. 2023. All things considered: Detecting partisan events from news media with cross-article comparison.

Xin Luo and Chu-Ren Huang. 2022. Gain-framed buying or loss-framed selling? the analysis of near synonyms in Mandarin in prospect theory. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 447–454, Manila, Philippines. Association for Computational Linguistics.

Yiwei Luo, Kristina Gligorić, and Dan Jurafsky. 2024. Othering and low status framing of immigrant cuisines in us restaurant reviews and large language models.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. *Theory and Applications of Categories*.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1438–1448, Beijing, China. Association for Computational Linguistics.

Tom Nicholls and Pepper Culpepper. 2020. Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 38:1–23.

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.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jinsheng Pan, Zichen Wang, Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2023. Understanding divergent framing of the supreme court controversies: Social media vs. news outlets.

Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5209–5235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. pages 3001–3022.

Marten Postma, Levi Remijnse, Filip Ilievski, Antske Fokkens, Sam Titarsolej, and Piek Vossen. 2020. Combining conceptual and referential annotation to study variation in framing. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 31–40, Marseille, France. European Language Resources Association.

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.

Carley Reardon, Sejin Paik, Ge Gao, Meet Parekh, Yanling Zhao, Lei Guo, Margrit Betke, and Derry Tanti Wijaya. 2022. BU-NEmo: an affective dataset of gun violence news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2507–2516, Marseille, France. European Language Resources Association.

Javier Sánchez-Junquera, Berta Chulvi, Paolo Rosso, and Simone Paolo Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11:3610.

Hernan Sarmiento, Felipe Bravo-Marquez, Eduardo Graells-Garrido, and Barbara Poblete. 2022. Identifying and characterizing new expressions of community framing during polarization.

Holli Semetko and Patti Valkenburg. 2000. Framing european politics: A content analysis of press and television news. *Journal of Communication*, 50:93 – 109.

Karthik Sheshadri, Chaitanya Shivade, and Munindar Singh. 2021. Detecting framing changes in topical news. *IEEE Transactions on Computational Social Systems*, PP:1–12.

Qi Yu. 2023. Towards a more in-depth detection of political framing. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 162–174, Dubrovnik, Croatia. 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, Seattle, Washington. Association for Computational Linguistics.

Jin Zhao, Jingxuan Tu, Han Du, and Nianwen Xue. 2024. Media attitude detection via framing analysis with events and their relations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17197–17210, Miami, Florida, USA. Association for Computational Linguistics.

Jin Zhao, Jingxuan Tu, Bingyang Ye, Xinrui Hu, Nianwen Xue, and James Pustejovsky. 2025. Beyond benchmarks: Building a richer cross-document event coreference dataset with decontextualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3499–3513, Albuquerque, New Mexico. Association for Computational Linguistics.

Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Baseline Model Selection

Since FRECO is a newly introduced task with no existing benchmarks, we establish a set of representative baselines using fine-tuned cross-encoders LLMs. We adopt SFT and DPO as our primary training strategies. While prior work has explored more complex methods such as PPO for preference modeling, we chose DPO due to its efficiency and competitive performance in practice. PPO requires significantly more computational overhead for policy learning and reward modeling, which makes it less feasible for our iterative mining pipeline and large-scale experiments. Our selected baselines strike a practical balance between modeling strength and scalability, and provide a foundation for future comparison on this task.

A.2 Finetuning Details

We run experiments across three random seeds (3407, 521, 108). We finetune `Llama-3.2-3B`[3] and `Llama-3.1-8B`[4]. We incorporate FlashAttention-2 to improve memory efficiency, 4-bit quantization to reduce memory overhead, and sequence length scaling (up to 4,096 tokens) to enable models to process long contexts. Fine-tuning is conducted using Unsloth (Daniel Han and team, 2023). We apply LoRA with rank 8, an α scaling factor of 16, and dropout of 0.05. Training is performed on GPU-accelerated hardware, with per-device batch sizes of 2, gradient accumulation steps of 4, and AdamW 8-bit optimization. The learning rate is set to 5e-6, and training runs for 6 epochs with a linear learning rate schedule. Each training session takes 1-2.5 hours. In order to remove the confounding variable of "more training" when comparing methods, we want to make sure each method (e.g., SFT alone, DPO alone, SFT→DPO, etc.) uses the same total number of training epochs. We Equalize the Total Training Budget: SFT→DPO and DPO→SFT involves 3 epochs SFT + 3 epochs DPO = 6 total, SFT-only and DPO-only were trained 6 epochs. All experiments are conducted on a single 40GB Tesla V100 GPU.

[3] https://huggingface.co/unsloth/Llama-3.2-3B-Instruct

[4] https://huggingface.co/unsloth/Meta-Llama-3.1-8B

A.3 Finetuning Prompts

We provide the fine-tuning prompts used in our experiments in Table 2. These prompts guide the model in learning to identify equivalent events with contrastive framing and classify individual event attitudes.

A.4 FRECO Corpus Annotation Guidelines

A.4.1 Introduction

This guideline provides instructions for annotating FRECO pairs in the FRECO corpus. The goal is to identify event pairs where actions, participants, locations, or context remain the same, but differences in description lead to distinct interpretations or emotional responses. These distinctions often result from word choices, connotation, and emphasis, shaping the perception of the event.

A.4.2 Definition of FRECO

FRECO pairs meet the following criteria:
Core Similarity:
(1) The events describe the same underlying action or situation.
(2) The events involve same or compatible participants, locations, or contextual references.
Framing Differences:
(1) Variations in word choice, syntax, or level of abstraction that shift interpretation.
(2) Differences in connotation (e.g., neutral vs. loaded terms).
(3) Emphasis on different causal interpretations or moral evaluations.

A.4.3 An Example

These two sentences in example 4 describe the same event, the Israeli military's actions at Shifa Hospital, but frame it differently, making them equivalent events with contrastive framing.

Sentence A refers to the event as a "raid at Gaza's largest hospital", a term that connotes force and aggression, while Sentence B describes it as an "operation in Shifa Hospital", which sounds more neutral and procedural. This lexical difference influences whether the action is perceived as an invasive assault or a strategic mission.

Additionally, Sentence A maintains neutrality by stating "where the military says Hamas was operating", presenting it as an assertion rather than fact. In contrast, Sentence B, spoken from the military's perspective, claims "we separated the patients and displaced civilians from the terrorists", framing

Setting	Prompt Instruction
SFT	`Determine if the event triggers tagged with _EVENT in the following two sentences refer to the same underlying event but with different framing. Answer with '1' if yes, '0' if no.` `[SentenceA, SentenceB]`
DPO	`In the following two sentences, the event trigger word is tagged with "_EVENT", Are the tagged two events equivalent but with different framing?` `[SentenceA, SentenceB]`

SFT - event attitude	`Determine whether the event triggers tagged with _EVENT in the following two sentences refer to the same underlying event but with different framing. Respond with '1' if they do and '0' if they do not.`
	
	`If they refer to the same event with different framing ('1'), also classify each tagged event's attitude toward the Al-Shifa hospital raid as supportive, neutral, or skeptical.`
	`[SentenceA, SentenceB]`

Figure 2: Finetuning prompts.

the action as a humanitarian effort rather than an aggressive attack. The use of "we" in Sentence B further reinforces an internal, justificatory framing, whereas Sentence A remains an external report.

These framing differences shift the reader's perception, shaping the event as either a military raid or a necessary operation, justifying its annotation as contrastively framed equivalent events in FRECO corpus.

(4) Sentence A: Israel 's army was on Thursday four days into a **raid** at Gaza 's largest hospital , where the military says Hamas was operating from among patients and displaced civilians .

Sentence B: During the **operation** in Shifa Hospital, we separated the patients and displaced civilians from the terrorists," he added .

A.4.4 Annotation Process

(1) Read both event descriptions in the containing sentence carefully.
(2) Determine if they describe the same underlying event. If not, label 0 for non-equivalent and discard.
(3) Identify the framing difference by examining participants, actions, location, context, or modality. See Table 3 for examples with different contrastive event components. If there's no framing difference, label 0 for non-equivalent and discard.
(4) Label event attitudes (supportive, neutral, or skeptical toward the main event).
(5) Provide justification for difficult cases in an annotation note.

Equivalent Event Pairs with Contrastive Framing	**Contrastive Event Components**
"**killing** two individuals" "**killing** two rioters"	**Contrastive Participants:** individual vs. rioters. Former is neutral, and the latter frames the people as disorderly.
"he was **hunted down** by Rittenhouse" "Rittenhouse was **pursuing** him"	**Contrastive Actions:** hunted down vs. pursuing. Former conveys a strong aggressive connotation, and the latter is neutral.
"the protest **took** place outside a government building" "the protest **happened** in a residential neighborhood"	**Contrastive Location:** outside a government building vs. in a residential neighborhood. Former implies a legitimate place for protest, and the latter frames the protest as disruptive to local residents.
"prosecution **provided** video" "the failure to **provide** video by prosecution"	**Modality Difference:** provided vs. failure to provide. Former presents the action affirmatively, and the latter indicates negligence.

Figure 3: Examples of FRECO. Each row illustrates how variations in participants, actions, locations, or modality lead to distinct event interpretations. Differences in word choice influence connotation, legitimacy, or emotional response—shaping the framing of the same underlying event.

A.4.5 Notes on Ambiguity

(1) If the action, participants, location, or context remain identical, and there is no contrast in framing, do not annotate as a FRECO pair.
(2) If an event pair is not coreferential but refers same underlying event or presents distinct frames, annotate based on framing contrast rather than strict event identity.
(3) Table 4 is designed to help annotators consider

event relations when determining whether an event pair qualifies as equivalent with contrastive framing. By distinguishing between different types of event relations—such as coreferential, concept-instance, whole-subevent, and superset-subset relations—annotators can assess whether two events describe the same underlying occurrence with differing frames rather than being strictly coreferential. Other near-coreference event relations as defined by O'Gorman et al. (2016) can also refer to the same underlying event. Unlike coreferential event pairs, which focus on identifying events that refer to the same real-world instance, our annotation task requires identifying cases where framing differences alter interpretation, connotation, or emphasis while maintaining semantic equivalence.

Equivalent Event Pairs with Contrastive Framing	Event Relations
"**killing** two individuals" "**killing** two rioters"	Coreferential
"**self-defense** in kenosha riots" "**fired** weapon in BLM protests"	Concept-Instance relation
"**de-Escalate** the situation" "**disperse** the unarmed protestors"	Whole-Subevent relation
"mass **layoffs**" "shooter **lost** his job"	Superset-Subset relation
"MSNBC **tells** stories about Rittenhouse" "narratives **peddled** by the talking heads of MSNBC"	Non-coreferential relation
"I was forced to **sell** my car to her due to financial hardship" "She **bought** my car"	Logically equivalent relation

Figure 4: Event Relations in FRECO. This table categorizes equivalent event pairs based on their event relations.

A.5 Downstream Evaluation Task: Media Attitude Detection

Framing plays a crucial role in shaping media attitudes, making Media Attitude Detection a natural downstream task for evaluating both FRECO corpus data quality and FRECO models. Responsible media outlets rarely fabricate facts outright; instead, their attitudes is often conveyed through framing — choosing specific angles, language, and contextual emphasis to subtly shape perception (Semetko and Valkenburg, 2000). Our FRECO model aims to capture such framings contrastively to help explain the attitudes of the news articles better.

Zhao et al. (2024) annotated a Media Attitude Detection (MAD) dataset by labeling each article as supportive, skeptical or neutral towards the topic main event. They encode three different framing device as model input, and finetune transformer models and prompting $\text{FlanT5}_{\text{XL}}$ and GPT-4o to label the attitude of each article.

Model	SFT	DPO→SFT
`Llama-3.2-3B`	88.43	89.26
`Llama-3.1-8B`	92.56	91.74
`Llama-3.2-3B` + SRL	90.08	89.26
`Llama-3.1-8B` + SRL	91.74	**93.39**

Table 4: Accuracy of attitudes classification in individual framing-divergent coreferential events.

A.5.1 Experiment Setup

We use the `Llama-3.1-8B` finetuned for the FRECO task to extract the FRECO from Zhao et al. (2024)'s MAD test dataset. Since the model extracts pairs of events rather than individual events, we separate the events within each pair and treat all extracted events as representative of their respective articles, so that each article is characterized by a set of framed events, which we call FRECO events for simplicity.

To demonstrate that FRECO events carry framing valence and can benefit MAD task, we encode them as a new contrastive framing device and evaluate their effectiveness. Following experimental setup in Zhao et al. (2024), we apply the $\text{FlanT5}_{\text{XL}}$ and GPT-4o zero-shot prompting to evaluate FRECO events device against their best performing setting in MAD task. We also use same number of randomly selected events from the test articles as a control to assess whether FRECO events provide a meaningful advantage over non-targeted event selection.

With the intuition that articles with more supportive events are likely to be supportive overall, we classify article-level attitudes by aggregating individual event-level attitudes using a majority voting strategy. In order to get the attitude label of individual FRECO events, we finetune `Llama-3.1-8B` on our FRECO corpus to jointly detect the FRECO and their individual attitude. The individual attitude detection results are shown in Table 4. If an event appears in multiple equivalent event pairs with conflicting labels, we resolve discrepancies by prioritizing supportive over neutral and skeptical over neutral. Notably, we found no instances where the same event was predicted

Topics	Framing Device	FlanT5$_{XL}$	GPT-4o	Majority Voting
PUTIN	Context events	70.69	81.38	N/A
	FRECO events	73.10	79.31	**82.07**
	Random events	37.93	46.21	N/A
AL-SHIFA	Context events	73.89	**80.00**	N/A
	FRECO events	74.48	77.24	74.48
	Random events	48.97	40.69	N/A
HONGKONG	Context events	65.46	78.17	N/A
	FRECO events	73.79	**79.31**	78.62
	Random events	46.21	53.10	N/A

Table 5: Comparison of Different Models and Majority Voting with different encoding of events. Context Events – The best-performing framing device reported from Zhao et al. (2024). Extracted Framing-Divergent Events – Events identified by our FEC model, which are expected to capture framing contrasts. Randomly Selected Events – A baseline consisting of an equal number of randomly sampled events from the test articles. Majority Voting - aggregated counts of attitudes of individual events identified by our FEC model, and only available for Framing-Divergent events

both supportive and skeptical in different extracted pairs, as framed events in this dataset are generally unambiguous. While there can be some ambiguity between neutral and either supportive or skeptical, direct contradictions between supportive and skeptical do not occur.

A.5.2 Results and Analysis

As shown in Table 5, FRECO events consistently improve attitude detection compared to random events, confirming their value as a framing device. GPT-4o generally outperforms FlanT5$_{XL}$, but the best-performing approach varies by topics. Majority voting with FRECO events achieves the highest score in one topic (Putin Election Win) and remains competitive in others, suggesting that aggregating individual event-level attitudes is an effective strategy. Context events remain a strong baseline, but FRECO events show comparable or better performance, indicating their effectiveness in capturing meaningful framing cues, highlighting the benefit of explicitly modeling contrastive framing.

Our FRECO events reduce input token counts by 62% comparing to originally article on average across all topics, achieving higher compression rates than other framing devices, making it more efficient for training. Additionally, FRECO events provide a more interpretable representation of an article's attitude.

Using FRECO-extracted events as framing devices allows us to more effectively capture an article's attitude toward a main event, such as Putin's election win. Because FrECo prioritizes framing-loaded event mentions. It surfaces events that are strategically positioned to shape reader interpretation. For example, a FRECO might extract an event like the assassination of an opposition leader, which implicitly critiques the legitimacy of the election, rather than more neutral mentions such as ballot counting. This makes FRECO a more targeted tool framing device extraction.

Framing differences, such as victimhood versus justification, skepticism versus endorsement, or causal attributions, are key mechanisms shaping media attitudes. Our results show that FRECO effectively capture these mechanisms and serve as strong indicators of media attitude. Evaluating our model on an attitude detection task demonstrates its ability to identify meaningful framing shifts. By identifying FRECO, our model enriches attitude classification by providing deeper event-level context and improving explainability in how different sources frame the same underlying events.