Computational Framing with Event-Based Methods

Wellesley College NLP Class

Jin Zhao

Brandeis University

Dec 5, 2025

What is Framing?

Two different News Headlines about the same climate protest event

- *Activists bravely protest climate inaction.*
- *Radicals disrupt city over climate agenda.*

Definition

Media framing is a powerful tool that shapes public perception by highlighting, omitting, or reinterpreting specific aspects of events.

Motivation of Event-Based Computational Framing Analysis

- While framing theory in communication studies emphasizes mechanisms such as selection, emphasis, and causal attribution, computational framing research today remains coarse-grained, relying on topic classification to approximate frames or isolated lexical cues and framing devices that fail to capture the deeper narrative structure.
- Events offer a portable unit of analysis across issues, languages and cultures. Event-based methods (context event, event coreference and event causal relations) moves framing research beyond surface-level proxies.

OpenFrames Prototype Demo

`https://openframes-app-fb4cc7ab1615.herokuapp.com/`

Chapter 1: Beyond Benchmarks: Building a Richer Cross-Document Event Coreference Dataset with Decontextualization

What is **C**ross-**D**ocument **E**vent **C**oreference (**CDEC**)

Definition

The task of detecting and linking event mentions across different documents that describe the same real-world occurrence.

Document 1: ...Anger **surged** after the jury's decision, and crowds gathered near the courthouse to express their dissent....

Document 2: ...Demonstrators assembled downtown, protesting what they described as a miscarriage of justice, ...and tensions **escalated** as chants echoed through the city....

Document 3: ...Public outrage over the ruling **intensified**, triggering large scale protests ...and drawing thousands into the streets to demand accountability....

Challenges: *Lexical variability* (surged/escalated/intensified) · *Cross-doc reasoning* (same participants? location? time?) · *Granularity* (assembled = protesting?) · *Framing differences*

Why is CDEC Annotation Hard?

1. Context Understanding

Read full articles to understand each event — who's involved, when/where it happened, what occurred

2. Exhaustive Pairwise Comparison — $O(n^2)$

Compare every event mention across documents

1,000 mentions → ∼500,000 pairwise comparisons

3. Ambiguity Resolution

Resolve participants, time, location, and action — often not explicit in the text

Result

Slow, labor-intensive, cognitively demanding → **existing datasets are small and limited**

Motivation

Current CDEC Datasets Are Limited

- **Small & sparsely annotated** — In ECB+: 95% of pairs are non-coreferent, 88% of sentences have no annotated events, only ∼1.87 sentences/doc annotated
- **Artificial ambiguity** — e.g., "Lohan admitted to rehab" vs. "Reid admitted to rehab" — simplified and unrealistic

Our Goal

Build a **richer**, more **scalable**, and more **representative** CDEC dataset — one that reflects the real challenges of cross-document reasoning in the wild

Key Idea: Decontextualization

Traditional Approach

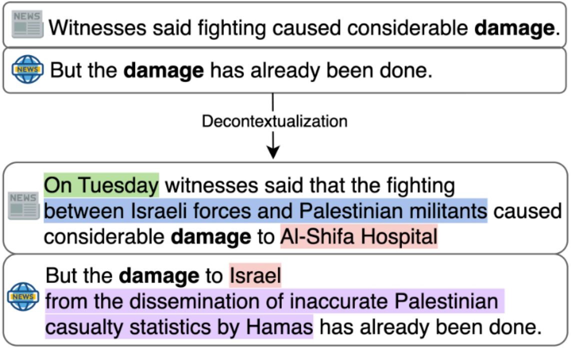
Read full documents, resolve coreference across large context

Our Approach

Annotate **sentence pairs** — inject context directly into each sentence using LLMs

Benefits:

- Adds explicit actors, locations, time, causes
- Sharpens coreference decision boundary
- Speeds up annotation significantly



Introducing **R**icher **E**vent**C**oref**B**ank (**RECB**)

What is RECB?

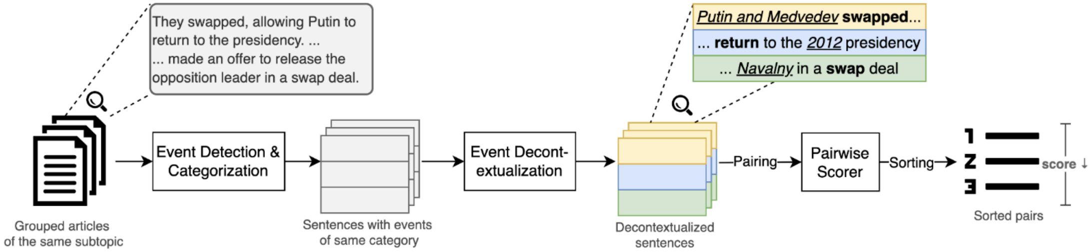
A CDEC dataset built entirely from **decontextualized sentences** — easier to annotate without sacrificing depth or realism

Advantages Over Traditional Datasets

- **Faster annotation** — sentences are self-contained
- **Higher density** — more coreferent pairs per annotation unit
- **Greater diversity** — diverse sources, richer event expressions

Document-Level Reconstruction

Each sentence maps back to its original document — preserving full context when needed



- **Data:** English news from 4 contentious topics, ideologically diverse outlets
- **Event Detection** → **Decontextualization** (o1-preview adds participants, time, location)
- **Scoring & Filtering:** BERT similarity + TF-IDF + verb constraints → avoids n^2 comparisons
- High-quality **ranked pairs** passed to annotators

Annotation

Task Design

Sentence-pair annotation using decontextualized event mentions

Label Types

Standard: `IDENTITY`, `NOT-RELATED`, `CANNOT-DECIDE`

Partial Coreference: `CONCEPT-INSTANCE`, `WHOLE-SUBEVENT`, `SET-MEMBER`

Procedure

- Progress through ranked pairs; stop after 200 consecutive `NOT-RELATED`
- 4 trained annotators, 400 pairs/subtopic double-annotated
- Joint "burn-in" phase + adjudication for disagreements

Quality

Cohen's $\kappa = 0.70$ (all labels) $\kappa = 0.78$ (binary)

Data Statistics

Topic	Source	Docs	Sentences	Ori / Decont. tokens	Mentions	Pairs	Near-Identity Pairs	Clusters
SHIFA	AAN	74	643	17k / 19k	1,267	6,834	406	353
	INN	58	692	17k / 20k	1,082	4,933	303	311
PUTIN	SN	77	1,047	29k / 32k	2,096	12,796	3,610	1,075
	GN	77	1,164	31k / 35k	2,346	12,197	3,690	1,094
HONGKONG	CD	76	868	22k / 26k	1,324	3,281	333	788
	GN	78	897	25k / 29k	1,677	5,226	368	1,046
RITTENHOUSE	TF	40	684	18k / 20k	1,025	1,679	364	493
	GN	64	1,340	34k / 36k	2,567	9,219	1,438	794
Total		**544**	**7,335**	**195k / 220k**	**13,384**	**51,665**	**10,512**	**5,954**

Table 2: Data statistics overview of RECB dataset. The number of articles, sentences, and tokens from each subtopic are reported after the data collection. We also report the number of event mentions, annotated pairs and cluster numbers from the human evaluation.

Comparison with Current Benchmarks

	RECB	ECB+	GVC
Docs	588	982	510
Sentences	7,335	15,812	9,782
Annot. sentences	7,121	1,840	4,604
Mentions	13,384	6,833	7,298
Clusters	5,954	2741	1,411
Non-singleton Clusters	2,358	1,958	1,048
Positive Pairs	26,756	26,712	50,799
Lemma-cluster Ratio	3.3	2.1	2.6
Cluster-lemma Ratio	5.6	3.5	2.0

Table 3: Comparison of the statistics on the RECB, ECB+, and GVC datasets.

RECB Advantages

- **4$\times$ more annotated sentences** than ECB+
- **2$\times$ mentions & clusters** — higher density
- **2,358 non-singleton clusters** — meaningful chains

Diversity Metrics

- **Lemma-cluster**: 3.3 (lexical diversity)
- **Cluster-lemma**: 5.6 (referential ambiguity)

RECB: richer annotations + greater linguistic complexity = more realistic benchmark

Experiments

Models

- **Lemma Matching** — links mentions with overlapping lemmatized surface forms
- **PairwiseRL** — fine-tuned RoBERTa cross-encoder for sentence pairs

Cross-Topic Evaluation (RECB)

Train on 3 topics, test on held-out 4th topic — evaluates **generalization across domains**

Cross-Dataset Comparison

- How do models perform on RECB vs. ECB+ and GVC?
- How well do RECB-trained models generalize to other benchmarks?

Lemma Matching Results

Test Split	CoNLL F1	Pairwise F1
ECB+	61.9	9.5
GVC	33.8	36.4
SHIFA	32.3	6.2
PUTIN	39.2	5.5
HONGKONG	48.2	**4.9**
RITTENHOUSE	**30.0**	5.9

Table 4: Lemma matching results on the test split of CDEC datasets. Pairwise F1 is based on the scores from all the sentence pairs, while CoNLL F1 is based on final

ECB+ (61.9 CoNLL F1)

High score due to **low lexical diversity** — many events use repeated surface forms

GVC (33.8 CoNLL F1, 36.4 Pairwise)

Some pairs share lemmas but don't form dense, transitive clusters

RECB (Lower Scores)

Reflects **intentional lexical diversity** — exactly the challenge needed for realistic CDEC progress

PairwiseRL Results

Train Split	Test Split (CoNLL F1)					
	ECB+	GVC	SHIFA	PUTIN	HONGKONG	RITTENHOUSE
ECB+	**82.9**	64.9	59.5	71.4	67.1	63.6
GVC	50.2	**84.4**	53.6	64.1	63.7	63.1
$RECB_{\text{-w/o Shifa}}$	80.2	62.9	**63.8**	-	-	-
$RECB_{\text{-w/o Putin}}$	82.4	64.8	-	**75.4**	-	-
$RECB_{\text{-w/o HongKong}}$	**82.9**	65.1	-	-	**68.3**	-
$RECB_{\text{-w/o Rittenhouse}}$	78.8	64.1	-	-	-	**68.5**

Table 5: Cross-evaluation results on the test split of CDEC datasets with pairwise-encoding.

In-Domain Performance

ECB+: 82.9 GVC: 84.4

Out-of-Domain Drops

ECB+ → GVC: 64.9
GVC → ECB+: **50.2** (overfit!)

RECB Generalization

- Cross-topic: solid 63–75 F1
- To ECB+: ∼**80+** (matches in-domain!)

Takeaway: RECB models are more robust due to richer lexical diversity

Conclusion

RECB: A New CDEC Dataset

High-quality, rich in diversity — built for realistic cross-document event coreference

Key Innovation: Decontextualization

Sentence-level, self-contained event mentions → **scalable, efficient, consistent** annotation without sacrificing realism

More Challenging Than ECB+ / GVC

Higher lexical variability, nuanced relations, fewer shortcuts for shallow models

Impact

A more realistic setting for developing **robust CDEC systems** — foundation for stronger generalization and deeper event understanding

Chapter 2: Media Attitude Detection via Framing Analysis with Events and their Relations

Introduction to Framing

What is Framing?

How media **highlights certain parts** of a story to shape a message or viewpoint (Entman, 1993)

Beyond Word Choices

Not just "protester" vs. "rioter" — we analyze how **events** are described, ordered, and linked

Events are the building blocks of narrative — how they're framed reveals the story's message

Our Goal

Not just to spot bias, but to help **understand and break down media narratives**

By learning framing strategies → think more critically about news attitudes

Event-Based Framing Devices

Device 1: Event Selection & Omission

What's included or left out changes the story

e.g., mentioning protests but not crackdowns — we group events to see what's emphasized

Device 2: Linguistic Framing

Word choice shapes perception: "protest" vs. "riot", "freedom fighter" vs. "terrorist"

We extract event triggers and arguments to capture this

Device 3: Causal Framing

Not just what happened, but **why** — one article credits a win to growth, another to suppression

We extract causal event pairs to map the narrative logic

Data Collection and Annotation

Three Polarizing Events

- Putin's re-election (March 2024)
- Al-Shifa Hospital raid (November 2023)
- Hong Kong July 1 protest (2019)

Big stories that different outlets frame very differently

Annotation Task

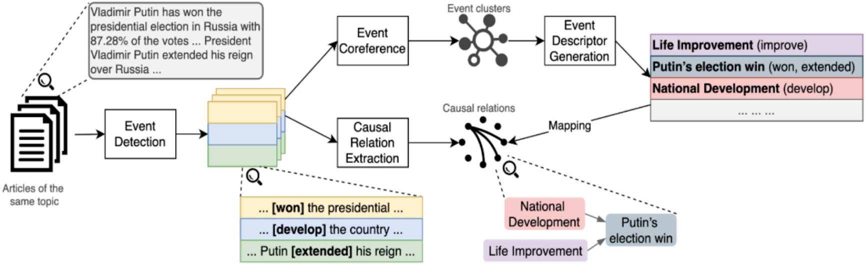
Label each article's **attitude** toward the main event:

- **Supportive** / **Skeptical** / **Neutral**

Counts	Putin	Al-Shifa	Hong Kong
Articles	495	643	471
Avg. tokens	314	232	297
Clusters	321	310	450
Avg. events	7	8	8
Avg. relations	7	9	9

Table 1: Statistics of the dataset for the media attitude task. The number of events and causal relations are reported after the filter.

Media Attitude Detection Pipeline



Pipeline: Articles → Event Extraction → CDEC → Shared Descriptors → Causal Links

Test with: Shared descriptors, original mentions, causal relations — which framing device best reveals attitude?

Example: Putin 2024 election win

...Vladimir Putin has **won** the presidential **election** in Russia with 87.28% of the votes after 100% of ballots were **counted**, the latest data from the Russian Central Election Commission (CEC) showed on Monday. Nikolay Kharitonov, the chairman of the lower house's Far East and Arctic Development Committee, **received** 4.31% of votes, while Leonid Slutsky, the chairman of the lower house's International Affairs Committee **got** 3.20%...Russians believe Putin is doing everything to **develop** the country and **improve** the lives of citizens...

...President Vladimir Putin **extended** his reign over Russia in a landslide **election** whose outcome was never in doubt, declaring his determination Monday to **advance** deeper into Ukraine and dangling new **threats** against the West.…Navalny **died** on February 16 in the Arctic prison where he was serving a 19-year sentence...Yevgeny Prigozhin, the head of the Wagner mercenary group with close ties to Putin, **died** in a plane **crash** with top associates. ...Sergei Yushenkov, a veteran politician and leader of the anti-Kremlin party Liberal Russia, is **shot** in front of his Moscow home...

Device 1: Selection and Omission of Events

Shared events:
Cluster1: Putin's election win - won, extended reign
Cluster2: Russian presidential election - election, election

Unique events in article 1:
Cluster3: completion of vote counting - counted
Cluster4: Kharitonov's Vote Share - received
Cluster5: Slutsky's Vote Share - got
Cluster6: National Development Efforts - develop
Cluster7: Life Improvements - improve

Unique events in article 2:
Cluster8: Military Advancement in Ukraine - advance
Cluster9: Threats to the West - threats
Cluster10: Navalny's Death - died
Cluster11: Prigozhin's Death - died
Cluster12: Plane Crash - crash
Cluster13: Yushenkov's Death - shot

Device 2: Linguistic Information

How Events Are Described

Captures how language shapes the story — trigger words (underlined) + arguments, location, time via SRL

Example: Same Event, Different Framing

Article 1: ...Vladimir Putin has <u>won</u> the presidential <u>election</u> in Russia ...

Article 2: ...President Vladimir Putin <u>extended</u> his reign over Russia in a landslide <u>election</u> whose outcome was never in doubt ...

"won" vs. "extended his reign" — same event, very different framing through word choice

Device 3: Cause and Effect Relations

Extracted explicit causal relations:
Cluster6 (National Development Efforts) → Cluster1 (Putin's election win)
Cluster7 (Life Improvements) → Cluster1 (Putin's election win)

Preconditions:
Cluster3: completion of vote counting → Cluster1 (Putin's election win)
Cluster4: Kharitonov's Vote Share → Cluster1 (Putin's election win)
Cluster5: Slutsky's Vote Share → Cluster1 (Putin's election win)

Context Events / implicit causal relations:
Cluster8: Military Advancement in Ukraine
Cluster9: Threats to the West
Cluster10: Navalny's Death
Cluster11: Prigozhin's Death
Cluster12: Plane Crash
Cluster13: Yushenkov's Death

Experiments Results (Fine-tuned Models)

		Fine-tuning		Prompting	
Topic	**Method**	**RoBERTa$_{BASE}$**	**T5$_{BASE}$**	**FlanT5$_{XL}$**	**GPT-4o**
Putin Election Win	Baseline	75.00	77.70	56.77	59.46
	Device 1	**82.07**	**83.45**	**70.69**	**81.38**
	Device 2	77.24	76.56	62.66	72.41
	Device 3	80.69	79.79	65.07	75.86
Al-Shifa Hospital Raid	Baseline	81.87	73.06	40.41	52.33
	Device 1	80.89	**75.56**	**73.89**	**80.00**
	Device 2	81.63	74.61	70.73	76.36
	Device 3	**82.25**	71.54	68.82	78.44
Hong Kong Protest	Baseline	**97.18**	**96.49**	53.50	52.52
	Device 1	93.02	91.80	**65.45**	**78.17**
	Device 2	93.71	87.43	60.45	73.54
	Device 3	94.41	89.97	63.32	77.48

Table 2: Evaluation results on the attitude detection task for each topic. We compare the baseline with inputs encoded from different devices. Accuracy from each model setting is reported.

Setup

Classification: RoBERTa (framing inputs)
Generation: T5 (QA-style prompts)
Baseline: Raw article text

Findings

- Topics vary in difficulty (Protest easiest)
- T5 doesn't consistently beat RoBERTa
- Framing inputs: **competitive** results
- **Shorter inputs** → more efficient training

Solid performance with compact, interpretable inputs

Experiments Results (Zero-Shot LLMs)

		Fine-tuning		Prompting	
Topic	**Method**	**RoBERTa$_{BASE}$**	**T5$_{BASE}$**	**FlanT5$_{XL}$**	**GPT-4o**
Putin Election Win	Baseline	75.00	77.70	56.77	59.46
	Device 1	**82.07**	**83.45**	**70.69**	**81.38**
	Device 2	77.24	76.56	62.66	72.41
	Device 3	80.69	79.79	65.07	75.86
Al-Shifa Hospital Raid	Baseline	81.87	73.06	40.41	52.33
	Device 1	80.89	**75.56**	**73.89**	**80.00**
	Device 2	81.63	74.61	70.73	76.36
	Device 3	**82.25**	71.54	68.82	78.44
Hong Kong Protest	Baseline	**97.18**	**96.49**	53.50	52.52
	Device 1	93.02	91.80	**65.45**	**78.17**
	Device 2	93.71	87.43	60.45	73.54
	Device 3	94.41	89.97	63.32	77.48

Table 2: Evaluation results on the attitude detection task for each topic. We compare the baseline with inputs encoded from different devices. Accuracy from each model setting is reported.

Models

Flan-T5 and GPT-4o in **zero-shot** setting

Raw Article Input

Performed **20+ points worse** than fine-tuned models

With Framing Inputs

Performance improved significantly!

Event descriptors + linguistic cues + causal links → **comparable to fine-tuned**

Takeaway: Even without fine-tuning, LLMs benefit from framing-aware inputs — *how we structure data matters*

Error Analysis

Model Input	Device	Label	Error type
Soldiers destroy hospital facility. ...	1	supportive / skeptical	CDEC error
Israel Defense Forces seizes weapons. ...	2	skeptical / supportive	SRL error
Implementation of extradition bill → restoration of order. ...	3	skeptical / supportive	Causal error

Table 4: Examples of common error types in the test set. We compare the predicted labels from GPT-4o with gold labels.

CDEC Errors

Wrong events grouped → misleading input

SRL Errors

Wrong agent/patient → bad attribution

Causal Errors

Missing causal links → incomplete signal

Insight: Most errors trace back to upstream extraction — pipeline quality is critical

How Framing Devices Help LLMs

Model Input	Device	GPT-4o Label
Navalny is murdered on February 16 in the Arctic prison where he was serving a 19-year sentence, Russia's prison service said...	Baseline	neutral
Navalny's death, ...	1	skeptical
Navalny is murdered in the Arctic prison ...	2	skeptical
Navalny's death $\rightarrow$ Putin's election win,...	3	skeptical

Table 5: Examples of GPT-4o with model input of baseline and different framing devices. We compare the predicted labels with gold labels.

Baseline

Raw article $\rightarrow$ neutral

Device 1

Event selection $\rightarrow$ skeptical

Device 2

Linguistic cues $\rightarrow$ skeptical

Device 3

Causal links $\rightarrow$ skeptical

Result: All three framing devices correctly identify stance that baseline misses

Conclusion

What We Showed

Framing-based approach reveals media attitudes by analyzing how events are **selected**, **described**, and **connected**

Why It Works

Models using structured framing inputs are:

- **Competitive** — comparable to fine-tuned models
- **Interpretable** — explainable event-based reasoning
- **Efficient** — concise, structured inputs

Main Challenge Ahead

Improving **coreference** and **causal extraction** quality — these upstream steps still limit overall accuracy

Chapter 3: Framing-Divergent Event Coreference for Computational Framing Analysis

What is **Fr**aming-divergent **E**vent **Co**reference (**FrECo**)

Same Event, Different Framing

Both sentences describe the **same real-world event**—an officer shooting someone—but frame it very differently

Positive/Justified Framing

Document 1: …The officer acted decisively to **neutralize the threat** ….

Critical/Negative Framing

Document 2: …The officer **opened fire** on the unarmed man ….

What FrECo Captures

These framing contrasts between **coreferential events**—same event, divergent perspectives

FrECo Task Definition

The Task

Finding pairs of event mentions that refer to the **same event** but are **framed differently**

Sources of Framing Divergence

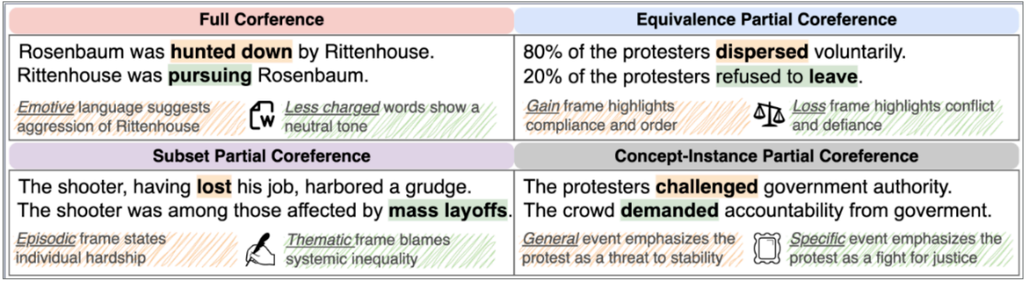
- **Word choice** — different lexical selections
- **Causal explanations** — different attributed causes
- **Emotional tone** — positive vs. negative valence
- **Narrative perspective** — different viewpoints or specificity

Two Formulations

- **Classification task** — given a pair, predict if it's FrECo
- **Mining task** — discover FrECo pairs at scale from large corpora

Building on CDEC Research

FrECo builds on the **relaxed identity** concept from event hoppers in CDEC research, incorporating both **fully** and **partially** coreferential event mentions



Annotation

Annotator Recruitment

- Two computational linguistics students
- Trained on definitions of FrECo and contrastive framing using detailed guidelines

Annotation Procedure

- Annotators label event mention pairs (ranked by CDEC similarity) as FrECo or not
- If FrECo, they also label each event's attitude toward the article's main event
- Joint review of 100 training pairs to align understanding

Agreement

Cohen's $\kappa = 0.76$ (FrECo identification)   Cohen's $\kappa = 0.81$ (attitude labeling)

FrECo Statistics

Total Data Size

3,800 annotated event mention pairs across 4 contentious news topics

Topic Breakdown

Putin: 739 pairs Al-Shifa: 1,356 pairs Hong Kong: 653 pairs Rittenhouse: 1,052 pairs

Label Distribution

- 1,765 pairs (46.5%) labeled as FrECo (framing-divergent coreferential)
- Remaining are non-coreferential or have no framing divergence

FrECo Pairwise Classification

Goal

Fine-tune classifiers to detect coreferent events with **divergent framing**

Two Input Variants

- **Raw context** — tagged event mentions in original text
- **SRL-enhanced** — highlights agents, patients, time, and location

Evaluation

Leave-One-Topic-Out Cross-Validation

4 topics: **Putin**, **Al-Shifa**, **Hong Kong**, **Rittenhouse**

Setup

- Train on 3 topics, test on held-out topic
- Dev set = 20% of train set (no test topic contamination)

Goal

Evaluate **generalization** across topics and framing strategies

Models

Baselines

- LLaMA-3.2-3B / 3.1-8B (zero-shot and fine-tuned)
- RoBERTa cross-encoder from prior CDEC work
- GPT-4 zero-shot

Fine-tuning Strategies

SFT, DPO, and combinations: **SFT→DPO** and **DPO→SFT**

Input Enhancement

SRL-enhanced inputs improve reasoning and structure awareness

Results Summary

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.

Error Analysis

False Negatives

Similar to regular CDEC errors:

- Context is too different between documents
- Miss partial coreferential cases

False Positives

Overgeneralization based on strong framing contrast alone

- Model sees framing divergence but events are not actually coreferent

Model Comparison

	PUTIN	AL-SHIFA	HONGKONG	RITTENHOUSE
$\texttt{RoBERTa}_{\texttt{BASE}}$	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.

Takeaway

RoBERTa baseline and GPT-4 zero-shot are **not as good** as fine-tuned LLaMA models

Bootstrapped FrECo Mining

Goal

Scale up from small annotated FrECo dataset

Approach

- Leverage gold-labeled pairs to mine high-confidence FrECo pairs from RECB corpus
- Use bootstrapping: iterative pseudo-labeling to expand coverage

Candidate Generation

Starting Point

Annotated FrECo pairs: **80% training**, **20% dev set** for validation

Scale Challenge

Full RECB corpus → ∼**4.87 million** candidate pairs (all events within each topic)

Filtering Strategy

- Use CDEC pairwise scorers to rank pairs by similarity
- Discard easy negatives with similarity < 0.3 (elbow point in distribution)
- Result: ∼**45K candidate pairs** remain

Final Pool

Includes original training data, excludes dev set and low-similarity tail

Bootstrapping Results

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.

Why Stop at Round 3?

- New positives drop sharply
- Val loss plateaus, then increases (Round 5)
- Jaccard similarity decreases → unreliable regions
- Manual review: noisy pairs dominate

By Round 3

- **6,693** new positive FrECo pairs mined
- Estimated **88% recall**
- Estimated **70.5% precision** (human eval)

Conclusions

New Task

Introduced **FrECo**: detecting divergent framing of the same events across media

What We Built

- Diverse annotated corpus across 4 contentious topics
- Fine-tuned LLMs for FrECo classification

Scaling Up

Bootstrapped mining achieves **high precision** across domains

Impact

Enables **interpretable, large-scale** framing analysis grounded in events

Chapter 4: Framing-Aware News Comparison Web Platform

Features

Side-by-Side Comparison of News Articles

- Contextual event selection & omission – via Event Extraction & CDEC
- Framing-sensitive causal links – via causal modeling w/ framing attributes
- Contrastively framed equivalent events – via FrECo framework

Users Can Explore How Media Construct Narratives

- Through event inclusion/omission
- Through chains of framing-driven causality
- Through diverging depictions of shared events

Purpose and Evaluation

Purpose

Demonstration of technical capabilities of our event-based framing pipeline

Human-Centered Evaluation

- Framing extraction quality
- Alignment with human perception of media framing

Expected Impacts

Broader Impact

- Connects computational framing analysis with media literacy applications
- Makes abstract framing structures visible and explorable

Provides New Ground For

- User feedback loops
- Trustworthy model evaluation
- Public education on framing tactics in news

Conclusion

Three Event-Based Framing Strategies

- Context Event Selection
- Framing-Divergent Coreferential Events
- Causal Construal Variations

Contributions

- A unified, event-centric framing analysis pipeline with both theoretical rigor and practical utility for understanding media narratives
- Public datasets, modeling, and a web-based demo to facilitate further research

Questions