Reframing Responsibility: Framing-Aware Event Causality Identification

Jin Zhao and **Jiayi Yao** and **Xinrui Hu** and **Nianwen Xue**
Department of Computer Science
Brandeis University
Waltham, Massachusetts, USA
{jinzhao,jiayiyao,Xinruihu,xuen}@brandeis.edu

Abstract

Causal explanations in political narratives are often framed and contested. Different sources may explain the same event by assigning responsibility to different actors, expressing different levels of certainty. Standard Event Causality Identification (ECI) focuses on detecting causal links and does not capture these distinctions. We introduce **Fr**aming-Aware **E**vent **C**ausality **I**dentification (FrECI), a framing-aware extension of ECI that models causal explanations as structured claims including responsibility targets, evaluative framing, source type, and epistemic modality grounded in established framing theories. We construct a multilingual dataset aligned across English, Chinese, and Arabic narratives using shared event anchors. We evaluate FrECI using prompt-based large language model baselines and supervised neural models. Results show that prompt-based baselines struggle to recover complete framed causal claims, while joint supervised models perform substantially better. Finally, we demonstrate that FrECI enables quantitative analysis of divergent causal attribution across narratives. The dataset and code are publicly available.[1]

Figure 1: Framing-aware causal explanations for a shared anchor event. The same effect event (*Israeli airstrikes*) is causally explained in three ways, differing in responsibility targets, framing, source, and certainty.

1 Introduction

Causal explanations play a central role in political narratives. When texts describe contentious events, they rarely present causality in a purely neutral way. Instead, they attribute responsibility to specific actors, frame those actors evaluatively, invoke sources to legitimize claims, and signal varying degrees of epistemic certainty. As a result, different sources often offer distinct causal explanations for the same real-world event, even when they appear to describe the same facts.

Figure 1 illustrates this phenomenon for a single anchor event, *Israeli airstrikes*. Across three documents, the same effect event is causally explained in different ways. In the first, airstrikes are framed as a response to Hamas rocket fire, exonerating Israel while assigning blame to Hamas. In the second, the same airstrikes are attributed to Israel's long-standing occupation and blockade of Gaza, shifting blame toward Israel and attributing the claim to an opposing source (*Palestinian officials*). In the third, airstrikes are discussed through a third-party lens (*international observers*), with neutral epistemic stance and neutral framing. Although all three sentences describe the same event, they differ systematically in *who is held responsible*, *how responsibility is framed*, *who advances the causal claim*, and *how certain the claim is*.

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

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46956–46977

Existing NLP work on Event Causality Identification (ECI) is not designed to capture these distinctions. Prior ECI formulations primarily focus on identifying whether a causal relation holds between two events, typically treating causality as a binary relation detached from attribution, framing evaluation, or epistemic stance. This puts limits on the applicability of ECI to political discourse and framing analysis, where causal explanations are inherently framed and contested.

In this paper, we introduce **Fr**aming-Aware **E**vent **C**ausality **I**dentification (FRECI), a novel task that extends ECI to explicitly model the interpretive structure of causal explanations. Rather than predicting causal links alone, FRECI requires models to recover *framed causal claims* that jointly encode: (i) a directed causal relation between events, (ii) epistemic modality capturing the certainty of the claim, (iii) source type indicating who advances the explanation, and (iv) responsibility targets paired with evaluative framing effects such as blame or credit. This formulation allows models to distinguish between causally similar but rhetorically divergent explanations.

In communication research, causal attribution is not treated as a neutral description of events, but as a central mechanism of framing through which responsible target (Iyengar, 1993), moral evaluation (Entman, 1993), epistemic stance (Bloome and Talwalkar, 1997) and source legitimacy (Tuchman, 1972) are constructed. FRECI operationalizes this insight by modeling causal explanations as framed claims rather than factual relations.

To support this task, we construct a new multilingual dataset using a semi-automated pipeline that bootstraps from high-quality event and causal annotations in MAVEN-ERE (Wang et al., 2022) and expands them into a politically contentious setting. Starting from English Wikipedia articles, we align parallel Chinese and Arabic narratives through Cross-Document Event Coreference (CDEC), recover candidate causal relations using a hybrid Large Language Model (LLM)-assisted approach, and annotate fine-grained framing attributes through expert adjudication. The resulting dataset contains 2,203 framed causal relations grounded in shared real-world event anchors.

We evaluate FRECI under an anchor-based setting using prompt-based large language models, a supervised joint neural architecture, and a Supervised Fine-Tuning (SFT) generative LLM. Our results show that while LLMs capture aspects of causal reasoning, explicit joint supervision is crucial for reliably recovering complete framed causal claims, particularly for responsibility attribution and framing effects.

Finally, we demonstrate that FRECI enables a new class of analyses beyond extraction accuracy. Because framed causal claims are aligned across documents via coreference anchors, FRECI provides a principled basis for measuring how narratives propose competing causal explanations for the same real-world outcomes. Concretely, this supports a workflow in which a structured FRECI extractor is run over multilingual coverage of an event of interest (e.g., an armed conflict or geopolitical crisis), the resulting framed claims are grouped by shared anchor events, and divergence across sources is quantified along the dimensions of responsibility, framing, source, and certainty. We instantiate this workflow in §4 through a *causal fragmentation* analysis that summarizes, for each shared outcome, how strongly competing causes are advanced across narratives—turning FRECI outputs into a directly interpretable measure of contested causal attribution.

In summary, this paper makes three contributions: (1) we introduce FRECI, a framing-aware extension of ECI; (2) we construct a multilingual, anchor-aligned dataset for studying framed causal explanations in political narratives; and (3) we provide both modeling benchmarks and empirical analyses demonstrating how FRECI enables quantitative study of divergent causal attribution across narratives.

2 FRECI

2.1 Task Overview

We introduce FRECI, a task that extends standard ECI by requiring models not only to identify causal links between events, but also to characterize the sociopolitical interpretation of those links.

Given a document and a set of event mentions corresponding to real-world events of interest (anchors), the model is required to identify causal relations involving those anchors and to predict structured attribution and framing information associated with each causal explanation.

2.2 Task Definition

Given a document D and a set of query event mentions $M_Q \subset \mathcal{S}(D)$, where $\mathcal{S}(D)$ denotes the set of all event mentions in D of a set of real-world

event anchors A, the goal of FRECI is to predict a set of framed causal assertions expressed in the document.

A causal explanation may involve multiple responsibility targets. We therefore define the atomic unit of prediction as a **target-specific causal assertion**, represented as a tuple:

$$(e_c, e_e, s, a_s, t, f_t)$$

where:

- $e_c \in M_Q$ is the **cause event mention**,
- $e_e \in M_Q$ is the **effect event mention**,
- $s \in \mathcal{S}$ is the **source type** indicating who advances the causal claim (`Author`, `Target`, `Ally`, `Opponent`, `Third_Party`),
- $a_s \in \mathcal{A}$ is the **epistemic modality** capturing the certainty (`Full_Affirmative`, `Partial_Affirmative`, `Neutral`, `Partial_Negative`, `Full_Negative`) with which the causal relation $\langle e_c, e_e \rangle$ is asserted by the source s,
- $t \in \mathcal{T}$ is a **responsibility target** held responsible for the causal outcome,
- $f_t \in \mathcal{F}$ is the **framing effect** assigned to target t (`Credit`, `Blame`, `Undermine_Credit`, `Exonerate_Blame`, and `Framing_Neutral`).

For a given causal relation between events (e_c, e_e), multiple responsibility targets may be present. In such cases, the causal explanation is represented as a *set of tuples*, one per target:

$$\{(e_c, e_e, s, a_s, t_1, f_{t_1}), \ldots, (e_c, e_e, s, a_s, t_k, f_{t_k})\}$$

source type s and epistemic modality a_s are **claim-level attributes** shared across all targets associated with the same causal relation, while framing effects f_t are **target-specific**.

A model prediction is considered correct if it correctly identifies the causal event pair (e_c, e_e) and recovers the complete set of target-specific tuples with the correct source type, epistemic modality, responsibility targets, and framing effects.

3 Dataset Construction Pipeline

We construct our dataset using a semi-automated pipeline that bootstraps from high-quality event annotations in existing resources and systematically expands them into a multilingual, adversarial political setting. The core idea is to anchor annotation on reliable event and causal structures, while exposing them to divergent narrative framings across languages and media contexts. The dataset construction pipeline consists of four stages: (1) multilingual corpus sourcing and expansion, (2) event extraction and CDEC, (3) hybrid causal relation identification, and (4) fine-grained attributive annotation of causal links.

3.1 Corpus Sourcing and Cross-Lingual Expansion

We use MAVEN-ERE as the seed corpus. MAVEN-ERE is a large-scale English Wikipedia dataset with high-quality annotations of event mentions, event causal relations and event coreference. To target politically contentious events, we manually select 51 MAVEN-ERE training articles whose topics are likely to elicit divergent interpretations across linguistic and cultural communities. For each English seed article, we retrieve the corresponding articles from the Chinese and Arabic editions of Wikipedia (sourced directly in their original language), yielding parallel narratives of the same real-world events.

Because MAVEN-ERE consists of curated English excerpts while the original-language Chinese and Arabic articles are substantially longer, we segment the non-English texts into multiple documents and translate them from their source language into English using `gpt-4o`, with prompts designed to preserve original causal wording and rhetorical structure (see Appendix 6). All translations are validated by fluent annotators. This translate-then-annotate design enables structured annotation and modeling in a unified English interface, while preserving cross-lingual signals of framing divergence in causal attribution that originate in the source-language articles.

Translation is a deliberate methodological choice: FRECI targets *structural* properties of causal explanations—responsibility targets, polarity of framing, source attribution, and epistemic stance—rather than lexical variation across languages. To assess whether translation alters these structural signals, we conducted a targeted audit on a sample of original-language sentences and their English translations and observed no polarity reversals in framing effects (e.g., `Blame` flipping to `Credit`) or in epistemic stance. Any residual attenuation of language-specific cues would, if anything, make cross-lingual divergence a conservative rather than inflated estimate.

3.2 Event Extraction and CDEC

Event mentions enter the dataset through two distinct paths. For the English seed articles, MAVEN-ERE provides *gold* event mentions and causal relations. For the translated Chinese and Arabic documents, where no gold annotations exist, we *automatically* extract event mentions using OmniEvent (Peng et al., 2023). We then select MAVEN-ERE events that participate in annotated causal relations and use them as anchors for cross-lingual alignment.

Using a pairwise cross-encoder CDEC model (Yu et al., 2022), we link automatically extracted mentions in the translated documents to their corresponding MAVEN-ERE anchor events, forming cross-lingual CDEC clusters. These clusters connect mentions of the same real-world *cause* and *effect* events across articles and languages. The anchor-based formulation is a deliberate design choice: it enables controlled comparison of structured causal explanations of *shared* real-world events across narratives. Mentions extracted from non-English documents that do not align with any anchor cluster fall outside the defined scope of FRECI rather than being lost; future work may extend the schema beyond anchor-aligned events.

For each target event, the resulting structure captures multiple, potentially competing, causal attributions proposed by different sources. This alignment enables the dataset to quantify causal attribution divergence by comparing how identical outcomes are causally explained across narratives.

3.3 ECI

We identify causal relations using a hybrid approach that combines LLM-based generative ECI with human refinement. Given the extracted event mentions and full document context, we prompt `gpt-4o` to propose candidate cause–effect relations, enabling the recovery of long-range and implicitly framed causal links that are difficult to capture with trained pairwise classifiers alone (Cheng et al., 2025). The prompt is provided in Appendix 7.

The LLM outputs are treated as a high-recall candidate set rather than as final labels. Human annotators subsequently perform a two-pass refinement: in the first pass, they validate correct links, remove spurious ones, and resolve ambiguities; in a second pass, annotators read the full document and *add any missed causal relations* not surfaced by the LLM, explicitly addressing potential false negatives in the candidate set. This hybrid strategy balances scalability with annotation quality.

Statistic	English	Chinese†	Arabic†	Total
Topics	49+2‡	49	2	51
Documents	51	612	21	661
Tokens	22,098	406,911	10,986	439,995
Events	817	4,479	224	5520
Coref. Chains		–		756
Non-S. Chains		–		569
Causal Relations	845	1,244	114	2,203
Responsible Targets	1774	2736	265	4775

Table 1: Dataset statistics. † denotes translated data. ‡ indicates that 49 English topics align with Chinese, while 2 English topics align with Arabic (topics are listed in Appendix 5). Coreference Chains and Non-Singleton Coreference Chains cross-document only and therefore reported as corpus-level totals.

Framing		Source		Certainty	
Value	%	Value	%	Value	%
Blame	21.4	Author	92.3	Full Aff.	90.9
Credit	24.2	3rd Party	2.5	Partial Aff.	6.4
Undermine Credit	9.8	Target	1.8	Neutral	1.2
Exonerate Blame	13.9	Opponent	2.2	Partial Neg.	0.3
Neutral	30.7	Ally	1.2	Full Neg.	1.1

Table 2: Framing-aware causal attributes distribution.

3.4 Causal Framing Attributes Annotation

For each validated causal relation, annotators apply a structured annotation schema to capture how responsibility are framed within sociopolitical narratives. First, we use Semantic Role Labeling (SRL)[2] to automatically identify candidate agents associated with the cause and effect events. Annotators then validate or revise these agent annotations. When an event has no explicit agent and denotes a broad societal outcome (e.g., economic decline, public health), the target defaults to the `Incumbent`.

Second, for each identified target, annotators label the framing effect (e.g., `Blame`, `Undermine_Credit`). For modality annotation, we assist annotation with rule-based detectors that flag candidate quoted sources and epistemic certainty cues (e.g., "allegedly"). Annotators validate these suggestions and finalize source type labels (e.g., `Author`, `Ally`, `Opponent`) and certainty levels.

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

3.5 Annotation Process

We employ a multi-stage annotation pipeline that combines model-assisted pre-processing with structured human adjudication to ensure annotation quality while managing the complexity of end-to-end causal framing analysis (see Figure 3 in Appendix for annotation interface).

3.5.1 Annotator Recruitment and Training

We recruited three annotators with backgrounds in computational linguistics. Annotators completed a calibration phase using a detailed guideline (see Table 2 in Appendix). A pilot annotation of 40 documents was jointly reviewed to resolve disagreements and align interpretations.

3.5.2 Two-Pass Annotation Workflow

We adopt a verify-and-refine workflow. for the causal relation verification in the first phase, annotators reviewed LLM-proposed causal relations, validating correct links, rejecting false positives, and adding missed causal relations.

For framing attributes annotation in the second phase, for each validated causal relation, annotators verified target and source spans and labeled framing attributes (Framing Effect, Source Category, and Certainty). Annotators were instructed to rely strictly on text-internal attribution signals, except when determining `Ally/Opponent` relations.

3.5.3 Adjudication and Agreement

To ensure consistency, 20% of topics were double-annotated. Disagreements were resolved through expert adjudication by a senior annotator, and recurring ambiguities informed iterative guideline refinement.

Inter-annotator agreement was evaluated separately for causal link identification, responsibility target identification, and framing attribution. For causal link identification, we report pairwise link F1. For framing annotations, agreement is computed only on causal links accepted by both annotators. Responsibility targets are treated as a set of text spans associated with each causal link. Target identification agreement is measured using overlap-based F1 with normalized span matching. Given a matched causal link and responsibility target, annotators assign a target-specific framing effect. Agreement for the framing effect is computed per matched target using Cohen's κ. In contrast, source type and epistemic modality are treated as causal claim level attributes shared across all targets associated with the same causal link. Agreement for source and modality is therefore computed once per causal link using Cohen's κ.

Component	Unit	Agreement
Causal Link	Event Pair	0.65
Target Identification	Span Set	0.71
Framing Effect	Target-level	0.62
Source Type	Claim-level	0.74
Epistemic Modality	Claim-level	0.88

Table 3: Inter-annotator agreement for FRECI. Agreement is measured using pairwise link F1 for causal link identification, span-level F1 for responsibility target identification, and Cohen's κ for categorical framing, source, and epistemic modality annotations.

Table 3 shows consistent inter-annotator agreement across FRECI components. Lower agreement for framing effects primarily reflects boundary cases in which weak or implicit attribution—such as subtle blame, credit, or exoneration—is difficult to distinguish from neutral framing, rather than systematic annotation error.

Dataset Scale and Positioning. While FRECI is anchored on 51 politically contentious topics, the resulting corpus comprises 661 documents, 5,520 annotated event mentions, 2,203 framed causal relations, and 4,775 responsibility-target annotations across English, Chinese, and Arabic (Table 1). Crucially, each causal relation is annotated along multiple structured dimensions—source type, epistemic modality, responsibility targets, and target-specific framing effects—so the effective supervision per instance is substantially denser than in standard binary ECI benchmarks. We therefore position FRECI as a *high-quality structured benchmark* for framing-aware causal attribution rather than a large-scale weakly supervised corpus. Scaling structured framing annotation to larger collections of documents and topics is an important direction for future work. We evaluate FRECI under an anchor-based setting, where gold anchor event mentions are provided as input. Given a document and a set of anchor event mentions corresponding to real-world events of interest, models are required to identify framed causal relations among these anchors and predict all associated attribution attributes.

3.6 Experimental Setup

Data Split. To avoid topic and event leakage, we construct *topic-held-out* splits. All documents asso-

ciated with the same topic—including the English seed article and its translated Chinese and Arabic counterparts—are assigned to the same split, so no test-time anchor event, document, or coreference cluster is ever seen during training. Specifically, we partition the 51 topics into 36 topics for training, 5 for development, and 10 for testing (details in Appendix 5). This setup imposes *event-level* distribution shift rather than mere instance-level variation: at test time, models are evaluated on previously unseen political events and controversies, with disjoint anchor clusters and disjoint actor inventories. Performance gains under this protocol therefore reflect generalization of the structured FRECI schema to new events, rather than fitting to the topical or lexical distribution of training documents.

Candidate Anchor Pair Generation. For each document, models form candidate cause–effect pairs by pairing each anchor-linked event mention with all event mentions in the same document. To control computational complexity, we restrict candidates to ordered mention pairs whose two event mentions occur within a fixed sentence window. All models are evaluated on the same candidate set to ensure fair comparison.

3.7 Models

Prompt-based LLM Baselines. We implement prompt-based LLM baselines using `GPT-4o` in zero-shot and chain-of-thought (CoT) settings. The model is prompted to identify framed causal claims involving gold anchor events and to output structured predictions in a constrained JSON format. Following the FRECI schema, GPT-4o predicts causal links between anchor events, assigns claim-level attributes (source type and Epistemic Modality), and generates a set of responsibility targets with target-specific Framing Effects.

In the zero-shot setting, the model is provided only with a task description, label definitions, and output schema. In the CoT setting, the model is additionally instructed to perform step-by-step reasoning prior to producing the final structured output; however, only the final output is evaluated. Both settings use identical output schemas and evaluation protocols. Full prompt templates 8 and 9 are provided in Appendix.

Neural Joint FRECI Model. We implement a joint neural model based on a pretrained `RoBERTa` encoder that factorizes causal reasoning and framing attribution into task-specific heads while sharing a common encoder and being trained end-to-end.

Given a document and an anchor-linked event mention, together with a candidate event mention from the same document, the model encodes the document using a pretrained $\text{RoBERTa}_{\text{LARGE}}$ with explicit anchor markers. Contextual representations of the anchor mentions and the document are combined into a shared pair representation.

On top of this representation, the model jointly predicts: (i) a directional causal link between the event mention pair, (ii) claim-level attributes including source type and Epistemic Modality, and (iii) responsibility targets and their associated Framing Effects. Responsibility targets are predicted by classifying text spans of candidate agent identified via SRL, and framing effects are predicted conditionally for each selected target. All components are trained jointly with a multi-task objective, enabling structured information sharing across causal and framing decisions.

SFT LLM Baseline. We additionally report an open-model baseline obtained by SFT `LLaMA-3.1-8B-Instruct` with QLoRA to directly generate framed causal claims. The model is trained to map from a document and a set of gold anchor event mentions to a JSON-formatted output representing complete causal claims, including claim-level attributes and target-specific framing. This approach performs *implicit joint modeling*, as all components of FRECI are predicted simultaneously via sequence generation without explicit task-specific heads. All predictions are evaluated using the same metrics as other models.

3.8 Evaluation Metrics

Models predict a set of target-specific causal assertion tuples of the form $(e_c, e_e, s, a_s, t, f_t)$. Evaluation is performed at both the tuple level and the causal-relation level.

Causal Link Identification. We evaluate causal link detection independently of framing attributes by computing precision, recall, and F1 over directed anchor event pairs (e_c, e_e). A predicted link is considered correct if both event mentions and direction match a gold causal link.

Responsibility Target Identification. For correctly predicted causal links, responsibility targets

Model	Causal Link			Attribution Components				Full Claim
	P	R	F1	T	f	s	m	EM
GPT-4o (Zero-shot)	34.0	41.5	37.4	73.1	28.4	74.4	91.5	14.6
GPT-4o (CoT)	61.5	17.0	26.6	81.4	29.1	79.4	98.5	15.2
Joint FRECI (RoBERTa)	49.7	72.6	59.0	77.3	73.1	75.6	96.0	27.4
w/o joint training (pipeline)	52.8	66.2	58.7	78.0	71.6	78.6	94.9	26.0
SFT LLaMA-3.1-8B	47.7	69.3	56.5	80.8	61.2	72.3	96.0	26.7

Table 4: Anchor-based FRECI results on held-out test topics. Causal link performance is reported as precision (P), recall (R), and F1 over directed anchor event pairs. Responsibility Target (T) is evaluated using span-level micro-F1, conditioned on correct causal links. Framing Effect (f) is evaluated using macro-F1 over matched targets. source type (s) and Epistemic Modality (m) are evaluated using macro-F1 at the claim level. Full-Claim EM requires correct prediction of the causal link, complete target set, claim-level attributes, and all target-specific framing effects.

t are evaluated as a set of text spans. We report span-level micro-F1 using one-to-one matching with an IoU threshold of 0.5, where IoU is defined as the token-level intersection over union between predicted and gold target spans. Each matched (e_c, e_e, t) corresponds to a target-specific causal assertion tuple.

Framing Effect Prediction. Given a correctly predicted causal link and matched responsibility target, the framing effect f_t is evaluated at the tuple level using macro-F1 over matched (e_c, e_e, t) instances.

Claim-Level Attributes. source type s and epistemic modality a_s are evaluated once per causal relation (e_c, e_e), since they are shared across all target-specific tuples derived from the same causal explanation. We report macro-F1 over these claim-level attributes.

Full-Claim Exact Match (EM). We additionally report Full-Claim EM, a strict end-to-end metric under which a predicted causal explanation is counted as correct if and only if the complete set of target-specific causal assertion tuples $(e_c, e_e, s, a_s, t, f_t)$ associated with a causal relation is recovered exactly. This requires correct identification of the causal event pair, the shared source type and epistemic modality, the full responsibility target set, and the framing effect assigned to each target.

3.9 Results

Table 4 reports anchor-based FRECI results on held-out test topics. Overall, the results show that recovering complete framed causal claims is challenging, and that explicit joint modeling substantially outperforms prompt-based and generative baselines.

Causal Link Identification. Prompt-based `GPT-4o` exhibits a clear precision–recall trade-off. Zero-shot prompting yields moderate recall but low precision, while Chain-of-Thoughts (CoT) prompting substantially increases precision at the cost of recall, resulting in lower F1. In contrast, supervised models achieve much higher causal link F1, with the Joint FRECI model performing best (59.0), followed closely by the SFT LLaMA baseline (56.5). These results indicate the importance of task-specific supervision for robust causal link detection.

Attribution Components. Responsibility target identification achieves relatively high performance across models when conditioned on correct causal links. However, framing effect prediction remains the most difficult component. `GPT-4o` baselines perform poorly on framing effect ($f \approx 28$–29 macro-F1), whereas the Joint FRECI model achieves a substantial improvement (73.1), outperforming the SFT LLaMA baseline (61.2). Source type and epistemic modality achieve uniformly high scores across models, reflecting both their relative stability and strong label skew in the data, where most causal claims are attributed to the `Author` and expressed with `Full_Affirmative`.

End-to-End Evaluation. EM provides a strict end-to-end measure, requiring correct prediction of the causal link, all responsibility targets, and all framing attributes. As expected, EM scores are much lower than individual component scores. Prompt-based GPT-4o achieves low EM (14.6–15.2), while both supervised models substantially

improve performance. The Joint FRECI model achieves the highest EM (27.4), narrowly outperforming SFT LLaMA (26.7), indicating that explicit joint modeling yields more consistent structured predictions.

Joint vs. Pipeline Ablation. To isolate the contribution of joint training from the underlying structured factorization, we evaluate a pipeline variant of the Joint FRECI model that removes multi-task coupling while keeping the encoder, candidate generation, and prediction heads identical: causal links are predicted first, and attribute heads are subsequently applied to predicted positives without shared optimization. The pipeline variant achieves comparable causal link F1 (58.7 vs. 59.0) but slightly lower Full-Claim EM (26.0 vs. 27.4). This indicates that the structured factorization and supervised attribute heads account for the majority of the gains over prompting and SFT baselines, while joint optimization yields a further benefit by coordinating link and attribution decisions under the strict end-to-end metric. Importantly, this confirms that the improvements observed in the joint model are not artifacts of implementation details but stem from the structured supervision central to FRECI.

Together, these results show that recovering complete framed causal claims requires models to jointly capture causal structure and responsibility attribution. While LLMs capture aspects of causal reasoning under prompting, explicit joint supervision is crucial for achieving consistent end-to-end performance under the strict EM metric. A breakdown of Joint FRECI performance by source-language origin (English seed, zh→en, ar→en) shows broadly comparable causal link F1 and Full-Claim EM across origins, indicating that the translate-then-annotate pipeline does not introduce a systematic disadvantage for non-English source material; full numbers are reported in Appendix A.7.

4 Competing Causes for Shared Outcomes

FRECI enables quantitative analysis of how narratives propose competing causal explanations for the same real-world outcome. Because event mentions are linked into CDEC clusters, we analyze causal competition at the level of event clusters.

Let E denote an effect event cluster representing a shared outcome. We collect all annotated causal relations whose effect event belongs to E. Each causal relation proposes a cause event that belongs to some cause cluster C, inducing an empirical distribution over cause clusters:

$$p(C \mid E) = \frac{n(C, E)}{\sum_{C'} n(C', E)},$$

where $n(C, E)$ counts causal relations linking a cause from cluster C to an effect in cluster E. Each causal relation contributes one vote, independent of the number of responsibility targets.

Causal Fragmentation. We quantify the degree to which an outcome admits competing causal explanations using *causal fragmentation*, defined as the normalized entropy of the cause distribution:

$$\mathrm{Frag}(E) = \frac{H(p(C \mid E))}{\log |C(E)|},$$

where $H(\cdot)$ is Shannon entropy and $C(E)$ is the set of distinct cause clusters observed for E. $\mathrm{Frag}(E) \in [0, 1]$ is 0 when all explanations concentrate on a single cause cluster and increases as explanations spread more evenly across alternative causes.

Our dataset contains 569 non-singleton event clusters and 2,203 annotated causal relations. We compute causal fragmentation for effect clusters with at least two associated causal relations to exclude degenerate single-claim cases. Figure 1 illustrates this phenomenon: the same outcome (*Israeli airstrikes*) is explained by multiple distinct cause clusters (e.g., *rocket fire*, *occupation/blockade*, *regional developments*), resulting in high causal fragmentation. Across eligible effect clusters, the average causal fragmentation is 0.44, indicating that many outcomes are associated with multiple competing causal explanations rather than a single dominant cause.

5 Related Work

5.1 Event Causality Identification

ECI has been extensively studied as the task of detecting binary cause–effect relations between events in text. Widely used benchmarks such as CausalTimeBank (Mirza et al., 2014), the Event StoryLine Corpus (Caselli and Vossen, 2017), and MAVEN-ERE (Wang et al., 2022) provide annotations of event-level causality within documents. These datasets do not model who advances a causal explanation, how responsibility is framed, or how

certainty is expressed, and they largely assume a neutral, context-independent notion of causality.

Methodologically, early neural approaches framed ECI as pairwise classification using contextual encoders (Zhang et al., 2015; Kadowaki et al., 2019). Subsequent work introduced document-level modeling, constructing event graphs and applying structured reasoning with graph neural networks or transformers to capture long-range dependencies and global consistency (Gao et al., 2019; Chen et al., 2022, 2023). More recently, prompt-based and generative approaches using large language models have shown strong few-shot and zero-shot performance (Shen et al., 2022; Xiang et al., 2023; Man et al., 2024). However, multiple studies report that LLM-based methods are sensitive to prompt design and prone to over-attributing causality, particularly in long or ambiguous contexts (Gao et al., 2023; Kıcıman et al., 2024). As a result, recent surveys and evaluations suggest that supervised models with explicit event representations remain the most reliable approach for ECI, with LLMs best used as auxiliary components (Wang et al., 2022; Chen et al., 2023; Cai et al., 2025).

FRECI builds on this line of work but departs from standard ECI by treating causal explanations as *framed claims*. In addition to identifying causal links, FRECI models responsibility attribution, evaluative framing, source type, and epistemic modality—dimensions not represented in existing ECI datasets or models.

5.2 Computational Framing

Computational approaches to framing have primarily focused on identifying issue- or topic-level frames in news and political discourse. Prominent datasets include 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), and issue-specific resources such as the Gun Violence Frame Corpus (Liu et al., 2019), VoynaSlov (Park et al., 2022), and stereoimmigrants (Sánchez-Junquera et al., 2021). Related work also addresses media attitude and bias detection (Zhao et al., 2024; Hamborg et al., 2019). These datasets have supported a range of computational methods, including topic modeling (DiMaggio et al., 2013; Nguyen et al., 2015), unsupervised learning (Burscher et al., 2016), semantic parsing (Ziems and Yang, 2021), and fine-tuned language models (Mendelsohn et al., 2021).

More recent work adopts an event-centric perspective on framing. Liu et al. (2023) aligns news articles covering the same story to analyze partisan event selection, while Das et al. (2024) clusters event relations into narrative structures. Zhao et al. (2024) examines contextual events surrounding a main event to identify framing through selection and omission. While these approaches move framing analysis closer to event structure, they do not explicitly extract causal claims of the form "X caused Y", nor do they model whether such claims assign blame, credit, or justification.

5.3 Cross-Lingual Narrative Analysis

Our work is also related to studies of narrative divergence across languages. Prior research using multilingual Wikipedia and parallel news corpora has shown that articles describing the same historical or political events often reflect language-specific or community-specific perspectives (Hecht and Gergle, 2010; Bao et al., 2012; Hale, 2014). Despite efforts such as Wikipedia's Neutral Point of View policy, cross-lingual analyses consistently find differences in emphasis, interpretation, and framing.

However, existing cross-lingual studies typically focus on high-level content differences, such as fact selection, topic emphasis (Samoilenko et al., 2017), or sentiment (Steinberger et al., 2011), rather than aligning specific causal explanations. FRECI enables the first structured analysis of divergent causal explanations for the same events across sources.

6 Conclusion

We introduced FRECI, a framing-aware event causality task that models not only causal relations, but how they are attributed and framed in political narratives. By anchoring explanations on shared real-world events, FRECI enables controlled cross-source and cross-lingual comparison of divergent causal attributions. Experiments show that recovering complete framed causal claims is challenging, and that joint supervision clearly outperforms prompt-based baselines. Overall, FRECI provides both a challenging NLP benchmark and a practical framework for analyzing framed causality in political discourse.

Limitations

Our study focuses on Wikipedia articles, which differ from time-sensitive news reporting in both authorship and editorial process. Consequently, the causal narratives captured in our dataset reflect stabilized, negotiated representations rather than immediate journalistic reactions. This makes Wikipedia a conservative testbed for studying framing: although a Neutral Point of View policy is enforced, neutrality is negotiated separately within each language edition, allowing framing differences to persist despite editorial normalization. As a result, our findings may not directly generalize to real-time media dynamics.

Our framing and source annotations rely on a discrete label set that enables clear operationalization and evaluation. However, real-world causal explanations may exhibit mixed, graded, or evolving framing that is not fully captured by categorical labels. Future work could explore continuous or probabilistic representations of causal responsibility and framing.

We evaluate FRECI under an *anchor-based* setting, where gold anchor event mentions are provided as input, isolating framed causal reasoning from upstream extraction and alignment errors. FrECI adopts an anchor-based formulation to enable controlled comparison of causal explanations across documents while avoiding unrestricted event pairing, making it well suited for cross-source and cross-lingual analysis. End-to-end and cross-narrative settings introduce additional challenges and are left for future work.

Because non-English articles are translated into English for annotation, some language-specific framing cues may be altered or attenuated despite human validation and our targeted audit (§3.1), which found no polarity reversals in framing effects. As a result, certain cross-lingual framing differences may be understated or over-regularized; observed cross-lingual divergence should therefore be interpreted as a conservative estimate.

Our dataset covers three languages and does not capture framing variation in other linguistic or geopolitical contexts. Extending FRECI to additional languages remains future work.

Finally, the dataset is intentionally curated rather than web-scale: 51 politically contentious topics yield 661 documents and 2,203 framed causal relations, with each instance carrying dense multi-dimensional annotation. This positioning supports controlled, high-quality evaluation of framing-aware causal attribution but may limit the training of data-hungry models. Scaling structured framing annotation to larger corpora is left for future work.

Ethical Considerations

This work analyzes political narratives and causal explanations drawn from publicly available Wikipedia articles. The dataset contains no private, personal, or sensitive individual-level information, and all source texts are released under Wikipedia's licensing terms. As such, the work poses minimal risk to individual privacy.

Our annotations model how causal explanations assign responsibility and evaluative framing within text. These labels are intended to describe narrative structure rather than to endorse any particular political interpretation. Nevertheless, framing annotations necessarily involve interpretive judgment. To mitigate annotator bias, we provide detailed annotation guidelines, conduct calibration and adjudication, and restrict annotators to labeling text-internal signals rather than external beliefs.

The dataset reflects the perspectives and norms of Wikipedia editor communities, which may not represent broader public discourse. As a result, analyses derived from this resource should not be interpreted as measuring societal consensus or ground truth about political events. Users of the dataset should exercise caution when drawing normative or policy conclusions.

Finally, while the proposed task and models can support analysis of political framing, they could also be misused to amplify or target particular narratives. We release this dataset and framework for research purposes and encourage responsible use, transparency, and critical interpretation when applying FRECI to real-world political content.

Acknowledgements

This work is supported by grants from the CNS Division of National Science Foundation (Awards no: NSF_2213804) entitled "Building a Broad Infrastructure for Uniform Meaning Representations". Any opinions, findings, conclusions or recommendations expressed in this material do not necessarily reflect the views of NSF. We also wish to extend our appreciation to Cloudbank, which provided an indispensable computational resource for our experiments.

References

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

Patti Bao, Brent J. Hecht, Samuel Carton, Mahmood Quaderi, Michael S. Horn, and Darren Gergle. 2012. Omnipedia: bridging the wikipedia language gap. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

David Bloome and Susan Talwalkar. 1997. Book reviews: Critical discourse analysis and the study of reading and writing. *Reading Research Quarterly*, 32(1):104–112.

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.

Ruichu Cai, Shengyin Yu, Jiahao Zhang, Wei Chen, Boyan Xu, and Keli Zhang. 2025. Dr.ECI: Infusing large language models with causal knowledge for decomposed reasoning in event causality identification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9346–9375, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. CHEER: Centrality-aware high-order event reasoning network for document-level event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816, Toronto, Canada. Association for Computational Linguistics.

Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects.

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.

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

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Scott A. Hale. 2014. Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, WebSci '14, page 99–108. ACM.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated identification of media bias by word choice and labeling in news articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Brent Hecht and Darren Gergle. 2010. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 291–300. ACM.

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

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality.

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.

Hieu Man, Chien Van Nguyen, Nghia Trung Ngo, Linh Ngo, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Hierarchical selection of important context for generative event causality identification with optimal transports. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8122–8132, Torino, Italia. ELRA and ICCL.

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.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

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.

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.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of ACL 2023*.

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.

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.

Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. 2017. Analysing timelines of national histories across wikipedia editions: A comparative computational approach. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, pages 210–219.

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.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Josef Steinberger, Polina Lenkova, Mijail Kabadjov, Ralf Steinberger, and Erik van der Goot. 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of Recent Advances in Natural Language Processing*, pages 770–775.

Gaye Tuchman. 1972. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of Sociology*, 77(4):660–679.

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

Wei Xiang, Chuanhong Zhan, and Bang Wang. 2023. Daprompt: Deterministic assumption prompt learning for event causality identification.

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.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China.

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.

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 FRECI Annotation Guidelines

See Figure 2

A.2 Annotation and Analysis Interfaces

To support consistent and efficient annotation of framed causal claims, we developed a web-based interface that presents anchor events, candidate causal relations, and structured framing attributes in a unified view. The interface enforces the FRECI schema and reduces annotator error by constraining label choices and maintaining shared claim-level attributes. See Figure 3 We additionally built a visualization interface to support qualitative analysis of framing-divergent causal explanations across aligned narratives. See Figure 4

A.3 Data Split

See Table 5

A.4 Prompt Templates for Translation

See Table 6

A.5 Prompt Templates for ECI

See Table 7

A.6 Prompt Templates for Prompt-based LLM Baselines

See zero-shot prompt at Table 8. See CoT prompt at Table 9

A.7 Cross-Lingual Performance Breakdown

Table 10 reports Joint FRECI performance on held-out test documents, grouped by source language: English seed articles, Chinese articles translated into English (zh→en), and Arabic articles translated into English (ar→en). Although our modeling operates in a unified English representation space after translation, the parallel construction of FRECI enables analysis of performance by document origin. Performance is broadly comparable across origins on both causal link F1 and Full-Claim EM, suggesting that the structured FRECI signals modeled by the joint architecture transfer reliably across narratives derived from different language editions.

A.8 Training Details for Joint FRECI model.

The Joint FRECI model is initialized with a pre-trained $\text{RoBERTa}_{\text{LARGE}}$ and fine-tuned end-to-end on the FRECI task. Input documents are truncated to a maximum length of 512 tokens. Special marker tokens are inserted to indicate the spans of anchor event mentions.

Candidate responsibility targets are extracted using semantic role labeling and restricted to agentive argument roles.

The model is trained using a multi-task objective that jointly optimizes all prediction heads. Directional causal link prediction and claim-level attributes (source type and Epistemic Modality) are optimized using cross-entropy loss. Responsibility target selection is optimized using binary cross-entropy loss over candidate targets. Target-specific Framing Effects are optimized using cross-entropy loss conditioned on gold responsibility targets. Attribute losses are masked for non-causal event pairs.

We use the AdamW optimizer with a learning rate of 2×10^{-5}. Models are trained for up to 10 epochs with early stopping based on development set Full-Claim EM. Batch size is 8.

A.9 Training Details for SFT LLM Baseline

We fine-tune `LLaMA-3.1-8B-Instruct` using SFT with QLoRA to produce structured FRECI outputs. Each training instance consists of a single document paired with its gold anchor event mentions as input, and a JSON-formatted list of framed causal claims as output. The output schema matches the prompt-based LLM baselines, including causal links, claim-level attributes, and target-specific framing.

Model and Optimization. We apply 4-bit quantization with NF4 and train LoRA adapters on attention projection layers. Training is performed using the AdamW optimizer with a learning rate of 2×10^{-4}. Models are trained for up to 3 epochs with early stopping based on development set Full-Claim EM.

LoRA Configuration. We use LoRA rank $r = 32$, scaling factor $\alpha = 64$, and dropout of 0.05. The maximum input sequence length is set to 2048 tokens. Batch size is adjusted via gradient accumulation to achieve a global batch size between 64 and 128.

Decoding and Output Constraints. At inference time, decoding is performed with greedy or low-temperature sampling. The model is instructed to output valid JSON only. Generated outputs are

Split	Topics
Train (n=36)	1953 Iranian Coup d'État, 2011 Chinese Pro-Democracy Protests, 2013 Tiananmen Square Attack, Annexation Of Crimea By The Russian Federation, Atomic Bombings Of Hiroshima And Nagasaki, British Expedition To Tibet, Caesar's Civil War, Cambodian-Vietnamese War, Finnish Civil War, Greek Civil War, History Of Xinjiang, Hungarian Revolution Of 1956, Indonesian Occupation Of East Timor, Iranian Revolution, Iran-Iraq War, Iraq War, Japanese Invasion Of Taiwan (1895), Korean Air Lines Flight 007, Lebanese Civil War, Libyan Civil War 2014-Present, Mukden Incident, Nepalese Civil War, Pacific War, Polish-Russian War Of 1792, Sino-Indian War, Sino-Vietnamese War, Somali Civil War, Soviet Invasion Of Poland, Spanish Civil War, Sri Lankan Civil War, Thallium Poisoning Case Of Zhu Ling, Turkish War Of Independence, War On Terror, Warsaw Pact Invasion Of Czechoslovakia, 2014 Gaza-Israel Conflict, Gaza War 2008-09
Dev (n=5)	2010 Hong Kong Electoral Reform, 2014 Pro-Russian Unrest In Ukraine, Korean War, Russo Japanese War, Slovak National Uprising
Test (n=10)	1963 Syrian Coup d'État, Air Raids On Japan, American Civil War, Angolan Civil War, Chinese Civil War, English Civil War, First Indochina War, Soviet-Afghan War, Warsaw Uprising (1794), Wukan Protests

Table 5: Train/Dev/Test Split of Topics

post-processed to remove extraneous text and validated for schema correctness prior to evaluation. Invalid or unparsable outputs are treated as empty predictions.

Evaluation Protocol. All outputs from the SFT LLM baseline are evaluated using the same evaluation protocol as other models, including causal link F1, attribution component scores, and Full-Claim EM. This ensures a fair comparison between implicit generative joint modeling and explicitly structured neural approaches.

Component	Instruction
System Role	You are a computational linguistics expert assisting in a study of **Adversarial Causal Attributions** in Wikipedia articles. You will translate text from `[SOURCE LANGUAGE]` to English.
Objective	The translation must transparently preserve the original author's causal logic, framing choices, and assignment of responsibility.
Causal Mapping	Every causal marker in the source text (connectors, resultative verbs, particles, or serial verb constructions) must have a 1:1 mapped equivalent in English. Do not paraphrase or restructure causality.
Agency Preservation	If the source uses passive voice to obscure or background an actor, retain the passive voice. If the source uses active or aggressive verbs to assign blame, do not soften or neutralize them.
Rhetorical Weight	Preserve the political or adversarial charge of terms. If the source uses a framing-loaded term (e.g., "martyred," "liberated," "betrayal"), select an English equivalent with the same framing intensity.
No Correction	Do not fix logical inconsistencies, balance viewpoints, or update facts to align with English Wikipedia. Faithfulness takes priority over accuracy or neutrality.
Language-Specific Instructions	
Chinese → English	Pay close attention to resultative verb compounds (e.g., , ,). Do not translate all as generic "caused." Use "resulted in" (often negative), "gave rise to," or "triggered" to match intensity. Preserve topic–comment structure and serial verb constructions when they imply causality without conjunctions. Maintain passive focus introduced by the *bei* () construction.
Arabic → English	Differentiate clause-based causality (*li-annahu*) from nominal causality (*bi-sababi*). If the particle *fa* implies immediate consequence, do not translate it as "and"; use "consequently," "thus," or "whereupon." Preserve metaphorical or abstract agency (e.g., attributing actions to concepts or forces) without sanitization.
Output	**[Translation]**

Table 6: Prompt for Analytically Faithful Translation in Adversarial Causal Attribution

Component	Instruction
System Role	You are an expert in Causal Inference and Natural Language Processing, specialized in event extraction.
Task	Identify directed cause–effect relationships between event mentions drawn only from a provided event list, using the document context.
Definitions	Direct Causality: Event A explicitly triggers Event B (e.g., "The bombing caused the shutdown.") Implicit / Long-range Causality: Event A creates necessary conditions, incentives, or motives for Event B, even when separated by paragraphs or mediated by discourse structure (e.g., a policy change early in the article leading to a protest later).

Input (User Template)	Document Context:
	`[INSERT FULL WIKIPEDIA ARTICLE TEXT]`
	
	Extracted Event Mentions:
	`[ID: E1, Text: "..."] [ID: E2, Text: "..."] ... [ID: En, Text: "..."]`

Output Format	Return a JSON list, where each item has the following schema:
	`{`
	`  "cause_id": "EX",`
	`  "effect_id": "EY",`
	`  "evidence_quote": "Exact sentence(s) justifying the link."`
	`}`

Table 7: Prompt for event-based causality extraction (ECI) from a document with a fixed event mention list.

Annotation Guidelines for Framing-Aware Event Causality Identification (FrECI)

Task	Given a document and a set of *anchor event mentions* corresponding to real-world events, annotate all **framed causal claims** involving these anchors. A framed causal claim specifies not only that one event causes another, but also how responsibility is assigned and framed in the narrative.
Causal Claim	A causal claim is a directed relation between two anchor events: • **Cause**: an anchor event presented as causally prior. • **Effect**: an anchor event presented as the outcome. Annotate if the text asserts or implies that the Cause brings about, explains, or leads to the Effect. Do **not** annotate purely temporal, correlational, or descriptive relations. **Hybrid annotation**: Review LLM-proposed causal relations, validate correct links, reject false positives, and add any missed relations.
Responsibility Targets	For each causal claim, identify all **responsibility targets**: actors, institutions, or groups portrayed as responsible for causing or enabling the effect. • Use SRL-extracted agents as candidates; validate or revise. • If responsibility is diffuse or no agent is named, annotate `Incumbent`.
Framing Effect	Assign a framing effect *independently* to each responsibility target: • `Blame`: target is negatively responsible or condemned. • `Credit`: target is positively responsible or praised. • `Undermine_Credit`: positive contribution is questioned or minimized. • `Exonerate_Blame`: target is justified or absolved of responsibility. • `Framing_Neutral`: responsibility stated without evaluation. When uncertain, default to `Framing_Neutral`.
Source Alignment	Assign **one** source label per causal claim (shared across all targets): • `Author`: asserted in the narrative voice (default if no attribution cue). • `Target`: attributed to a responsibility target. • `Ally`: attributed to a source aligned with the target. • `Opponent`: attributed to a source opposing the target. • `Third_Party`: attributed to external observers or analysts.
Epistemic Modality	Assign **one** modality per causal claim (shared across all targets): • `Full_Affirmative`: unhedged assertion. • `Partial_Affirmative`: hedged support (e.g., *likely*, *appears to*). • `Neutral`: reporting stance without commitment. • `Partial_Negative`: qualified doubt. • `Full_Negative`: explicit rejection of causality.
Annotation Scope	• **Claim-level**: Source Alignment and Epistemic Modality apply to the entire causal claim. • **Target-level**: Framing Effects are assigned independently to each responsibility target. • A single causal claim may have multiple targets with different framing effects.
Edge Cases	• If multiple causal claims appear, annotate each separately. • If multiple sources are cited, choose the source most directly asserting the claim. • If no causal relation exists between anchor events, annotate nothing. • Follow the text's assertion, not real-world plausibility.

Table: Human annotation guidelines for FrECI. Annotators identify framed causal claims between anchor events, annotate responsibility targets with framing effects, and assign claim-level source alignment and epistemic modality.

Figure 2: Annotation Guidelines for FrECI.

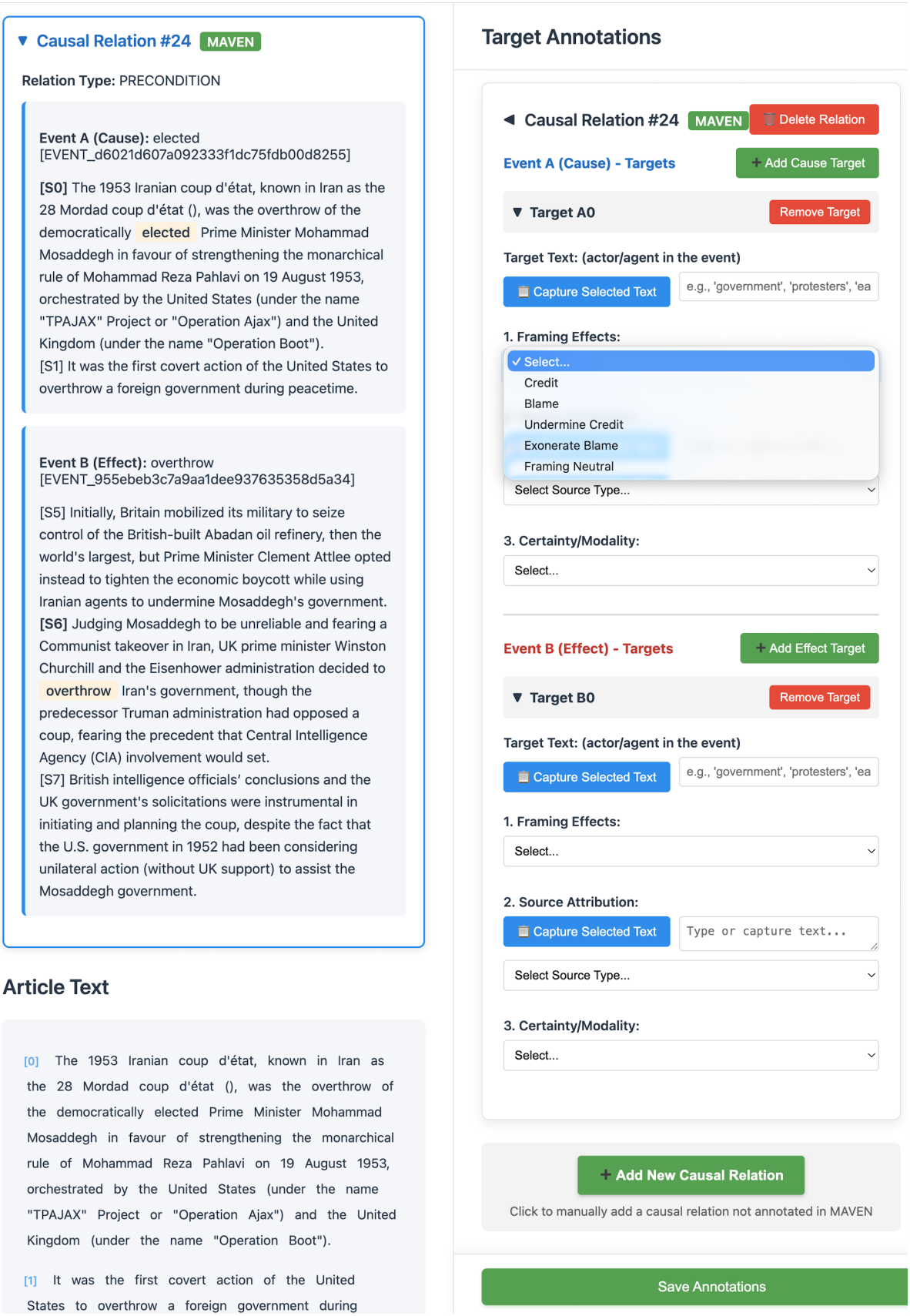

Figure 3: Web-based annotation interface used for FRECI. The interface presents gold anchor events, candidate causal relations, and structured annotation fields for responsibility targets, framing effects, source type, and epistemic modality. Claim-level attributes are shared across targets, while framing effects are assigned per target, enforcing the FRECI schema during annotation.

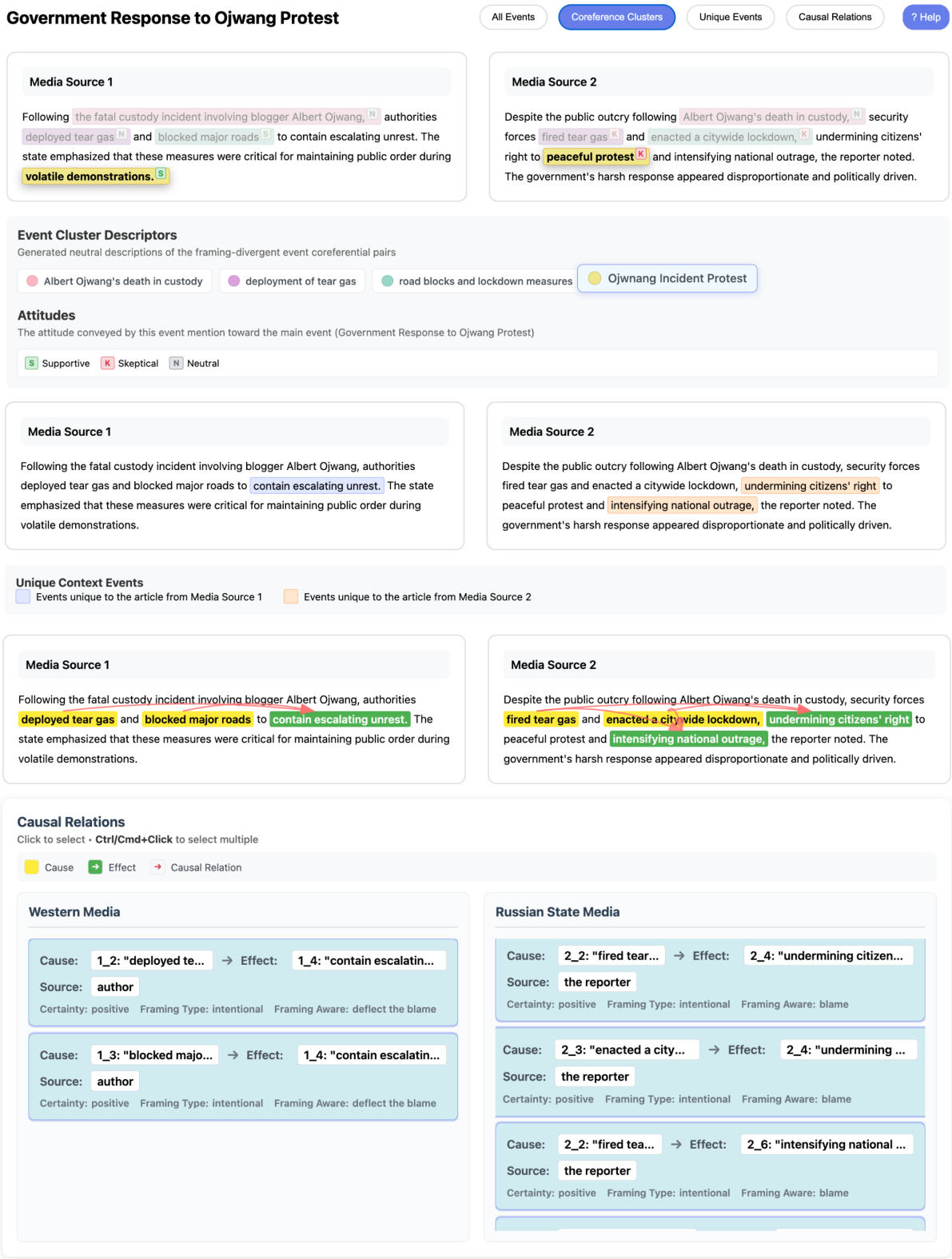

Figure 4: Visualization interface for analyzing framing-divergent causal explanations across aligned narratives. The interface highlights shared and unique events, causal relations, and responsibility attributions across sources, supporting qualitative analysis of divergent causal framing enabled by FRECI.

Zero-shot Prompt for FRECI

Task.

You are given a news document and a list of anchor event mentions corresponding to real-world events of interest. Your task is to identify *framed causal claims* involving these anchor events.

A framed causal claim consists of:

- **cause**: an anchor event that causes another anchor event
- **effect**: an anchor event that is caused
- **source**: who expresses the causal claim, one of {`Author`, `Target`, `Ally`, `Opponent`, `Third_Party`}
- **modality**: certainty of the causal claim, one of {`Full_Affirmative`, `Partial_Affirmative`, `Neutral`, `Partial_Negative`, `Full_Negative`}
- **targets**: a list of responsibility targets, where each target has:
  - **target**: the actor held responsible
  - **framing**: one of {`Credit`, `Blame`, `Undermine_Credit`, `Exonerate_Blame`, `Framing_Neutral`}

Rules.

- Use only the provided anchor events as causes or effects.
- Source and modality apply to the entire causal claim and are shared across all targets.
- Framing effects may differ across targets.
- If no causal relation exists between anchor events, output an empty list.
- Output *valid JSON only*. Do not include explanations or extra text.

Document:

`[DOCUMENT TEXT]`

Anchor Events:

`[ANCHOR EVENT LIST]`

Output Format (JSON):

```
[
  {
    "cause": "...",
    "effect": "...",
    "source": "Author|Target|Ally|Opponent|Third_Party",
    "modality":"Full_Affirmative|Partial_Affirmative|Neutral|Partial_Negative|Full_Negative",
    "targets": [
     {"target": "...", "framing": "Credit|Blame|Undermine_Credit|Exonerate_Blame|Framing_Neutral"}
    ]
  }
]
```

Table 8: Zero-shot prompt used for the prompt-based LLM baseline in FRECI.

Chain-of-Thought Prompt for FRECI

Task.
You are given a news document and a list of anchor event mentions corresponding to real-world events of interest. Your task is to identify *framed causal claims* involving these anchor events.

A framed causal claim consists of:

- **cause**: an anchor event that causes another anchor event
- **effect**: an anchor event that is caused
- **source**: who expresses the causal claim, one of {`Author`, `Target`, `Ally`, `Opponent`, `Third_Party`}
- **modality**: certainty of the causal claim, one of {`Full_Affirmative`, `Partial_Affirmative`, `Neutral`, `Partial_Negative`, `Full_Negative`}
- **targets**: a list of responsibility targets, where each target has:
  - **target**: the actor held responsible
  - **framing**: one of {`Credit`, `Blame`, `Undermine_Credit`, `Exonerate_Blame`, `Framing_Neutral`}

Instructions.
Before producing the final answer, reason step by step about:

- whether any anchor event causally leads to another anchor event;
- who is held responsible for the outcome;
- how responsibility is framed for each target;
- who expresses the causal claim and how certain it is.

After completing the reasoning, output *only* the final structured result in valid JSON.

Rules.

- Use only the provided anchor events as causes or effects.
- Source and modality apply to the entire causal claim and are shared across all targets.
- Framing effects may differ across targets.
- If no causal relation exists between anchor events, output an empty list.
- Do not include reasoning steps or explanations in the final output.

Document:
`[DOCUMENT TEXT]`

Anchor Events:
`[ANCHOR EVENT LIST]`

Output Format (JSON):

```
[
  {
    "cause": "...",
    "effect": "...",
    "source": "Author|Target|Ally|Opponent|Third_Party",
    "modality":"Full_Affirmative|Partial_Affirmative|Neutral|Partial_Negative|Full_Negative",
    "targets": [
     {"target": "...", "framing": "Credit|Blame|Undermine_Credit|Exonerate_Blame|Framing_Neutral"}
    ]
  }
]
```

Table 9: Chain-of-thought prompt used for the prompt-based LLM baseline in FRECI. The model is instructed to perform internal reasoning before producing the final structured output, though only the final output is evaluated.

Language Origin	Causal Link F1	Full-Claim EM
English (seed)	60.1	27.1
zh $\rightarrow$ en	58.3	27.4
ar $\rightarrow$ en	58.7	28.1

Table 10: Joint FRECI performance on held-out test documents, grouped by source language.