# Integrating Causal Reasoning into Automated Fact-Checking

Youssra Rebboud
youssra.rebboud@eurecom.fr
EURECOM
Sophia Antipolis, France

Pasquale Lisena
pasquale.lisena@eurecom.fr
EURECOM
Sophia Antipolis, France

Raphael Troncy
raphael.troncy@eurecom.fr
EURECOM
Sophia Antipolis, France

## Abstract

In fact-checking applications, a common reason to reject a claim is to detect the presence of erroneous cause-effect relationships between the events at play. However, current automated fact-checking methods lack dedicated causal-based reasoning, potentially missing a valuable opportunity for semantically rich explainability. To address this gap, we propose a methodology that combines event relation extraction, semantic similarity computation, and rule-based reasoning to detect logical inconsistencies between chains of events mentioned in a claim and in an evidence. Evaluated on two fact-checking datasets, this method establishes the first baseline for integrating fine-grained causal event relationships into fact-checking and enhance explainability of verdict prediction.

## CCS Concepts

• **Information systems** → **Decision support systems**; • **Computing methodologies** → **Information extraction**; **Causal reasoning and diagnostics**.

## Keywords

Fact-checking, Causal Reasoning, Explainability

**ACM Reference Format:**
Youssra Rebboud, Pasquale Lisena, and Raphael Troncy. 2026. Integrating Causal Reasoning into Automated Fact-Checking. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3748522.3779831

## 1 Introduction

While fact-checking spans a wide range of subtasks, a fundamental component is the assessment of entailment between a claim and its corresponding evidence. Existing models for textual entailment, often based on deep learning, are widely used for verdict prediction [5]. However, these models typically operate as black boxes, offering limited insight into their reasoning processes. To address this, prior efforts have explored explainability through attention mechanisms, summarization, or symbolic rule extraction [7].

Recent approaches rely on detecting cause-effect relationships between events described in the claim and those found in the evidence [3, 13, 14]. However, entailment between claim and evidence is not solely based on a general notion of causality, but also on a number of nuances of this concept. We can use an example:
**Claim:** *Taking the vaccine prevented infection.*
**Evidence:** *The vaccine triggered an immune response that blocked the virus from spreading in the body.*
Existing causal reasoning models typically represent relations only through general-purpose *cause* links, without distinguishing fine-grained relations such as *prevent*, *enable*, or *intend*. As such, they look for a (not existing) direct match: *Vaccine* $\xrightarrow{\text{cause}}$ *No infection*. Moreover, many systems oversimplify causal relations by skipping intermediate steps, obscuring the underlying logic to users. Considering the following notation:

$A$ : *Taking the vaccine* $C$ : *Immune response*
$B$ : *Infection* $D$ : *Virus blocked*

and the following relations:

- $A \xrightarrow{\text{causes}} C$ (*Vaccine causes immune response*)
- $C \xrightarrow{\text{prevents}} B$ (*Immune response prevents infection*)

an automated systems should infer:

$$A \xrightarrow{\text{causes}} C \xrightarrow{\text{prevents}} B \Rightarrow A \xrightarrow{\text{prevents}} B$$

In this work, we propose a causal explanation-based verdict prediction system that relies on semantically-precise event relations – namely *cause*, *prevent*, *intend* and *enable* – derived from the FARO ontology [9]. Our approach only applies to claims that include at least one causal relation between events. The use of semantically defined relationships ensures that the explanations align with human reasoning. In summary, this work makes the following contributions:

(1) we propose a set of high-level reasoning rules for verdict prediction to be applied to causal relations;
(2) we develop a complete pipeline for applying those rules to sentences in claim-evidence pairs.

All data, software, experiments, and a more detailed description of this work are publicly available at https://github.com/ANR-kFLOW/Fact_checking_reasoner.

## 2 Reasoning Rules

The FARO ontology [9] defines a set of semantically precise event relations. This ontology provides a textual definition to several event relations and defines logical axioms such as transitivity or disjunction using the OWL representation language. We focus on four relations: *direct-cause*, *prevents*, *intends-to-cause*, *enables*.

Throughout this section, we use four placeholders – $A$, $B$, $C$, and $D$ – to represent events (or entities) that can be related by *cause*, *enable*, *intend*, *prevent*, or *no-relation*. We consider the events $A$ and $B$ and their relationship "$A \rightarrow B$" in the claim and $C$ and $D$ with the relationship "$C \rightarrow D$" in the evidence. In the following subsections, we outline four key scenarios

**Logical alignment** is verified if the claim and the evidence include the same (or similar, or transitively-linked) events, which


*SAC '26, Thessaloniki, Greece*

ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3779831

are also connected by the same relation. The evidence *supports* the claim through logical alignment if the relation in claim and in evidence is the same and at least one of the following cases is verified: (1) $C$ is similar to $A$ and $D$ is similar to $B$; (2) a possible relation exists between $A$ and $C$ and/or between $B$ and $D$ which offer partial support by transitivity. In other words, while similarity between events provides a clear pathway to alignment, a direct causal connection can also strengthen the claim in cases where event similarity is not established.

**Logical Misalignment** is verified when the relation in the evidence and the one in the claim can be opposite. If we find a similarity matching ($C$ is similar to $A$ and $D$ is similar to $B$), we can conclude a direct contradiction to the claim: the same event cannot both cause (or enable/intend) and prevent the same outcome, making the evidence more likely to *refute* the claim.

**Causal loops** can be found among four events $A$, $B$, $C$, and $D$ by looking at the relationships (cause, enable, intend, or prevent) between each pair. We first take a claim ($A \rightarrow B$) and an evidence ($C \rightarrow D$) and infers how $A$ might relate to $C$ and how $D$ might relate to $B$. If all four relationships form a consistent cycle (such as a chain of causes, enables, or intends), we have a closed causal loop, which implies a high probability that the evidence is **supporting** the claim. Since "prevent" is by definition considered the cause of not happening of another event, two consecutive "prevent" relations effectively become a "cause" because of the effect of a double negation.

**Cherry-picking** is a term commonly used to define internal inconsistencies or selective usage of evidence. We group all evidence entries under the same claim, and then compares each pair of evidence elements. Each piece of evidence is represented as a $\langle$sub, rel, obj$\rangle$ triple, where "sub" and "obj" are events or entities, and "rel" is the relationship between them. The code measures how similar these events/entities are (e.g. $\text{sub}_1$ vs. $\text{sub}_2$, $\text{obj}_1$ vs. $\text{obj}_2$).

A claim is flagged for cherry-picking if certain patterns in the evidence emerge. For instance, it checks whether two pieces of evidence use the **same relationship** ($\text{rel}_1 = \text{rel}_2$) but involve subjects or objects that are dissimilar or opposites. If any of these mismatches is found, we deem the set of evidence potentially cherry-picked, because the evidence is either inconsistently presented or selectively used to reinforce the same relation in conflicting ways.

## 3 Methodology

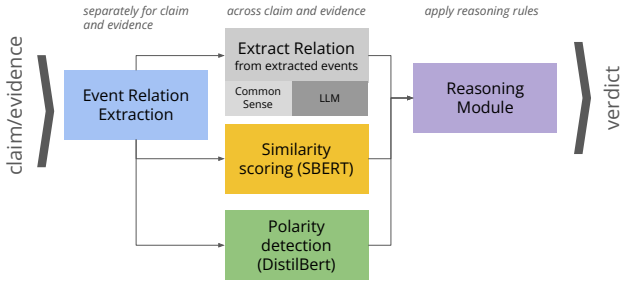


**Figure 1: Overview of our proposed pipeline**

Figure 1 illustrates our pipeline, whose steps are detailed in the following.

### 3.1 Causality Extraction within Claim/Evidence

It is performed within the same context (either the claim or the evidence) leveraging the sequence-to-sequence model REBEL [6], trained on a previously available annotated news dataset [8].

### 3.2 Causality Extraction across Claim/Evidence

We have experimented with two different strategies

*Common Sense-based Causality Extraction.* We employed the knowledge base ATOMIC [11], identifying the overlap with our studied causality relations.[1] We isolated the ATOMIC tuples involving these properties, creating a reference dataset. To address the absence of *enabling* and *prevention*, we have expanded the dataset by generating new examples using a LLM.

To include *no-relation* sentences, we use negative sampling on 50% of the dataset, stratifying on the relation type. Table 1 shows the final support of the common sense dataset, and the positive results obtained on the test set.

| Class | Support | Precision | Recall | F1-Score |
|---|---|---|---|---|
| `cause` | 82,242 | 0.8248 | 0.8424 | 0.8335 |
| `intend` | 146,588 | 0.8523 | 0.8924 | 0.8719 |
| `prevent` | 53,454 | 0.9849 | 0.9929 | 0.9889 |
| `enable` | 65,485 | 0.9755 | 0.9776 | 0.9765 |
| `no_relation` | 173,886 | 0.8669 | 0.8208 | 0.8432 |

**Table 1: Results of causality extraction between Claim Events and Evidence Events using ATOMIC augmented with LLMs**

*LLM-based Causality Extraction.* We experimented with the *Phi-3-Medium-4K-Instruct* model [1], using a prompt in two steps, first the classification a trivial relation – *(earthquake, death)→cause* –, then the actual annotation. We extracted and manually assessed 40 annotation, and out of the 40 samples, 33 were correctly processed (82.5%).

### 3.3 Similarity, Dissimilarity, and Opposites

Except for the exact match case, we rely on sentence similarity and dissimilarity, computed using SentenceBERT [10] on the concatenation of the event span and its original sentence. We empirically set the threshold to **0.54** to considered two events as similar.

Sometimes events represent concepts that are simply *dissimilar* (indicated by a similarity score falling below a certain threshold), while in other cases, they represent exact *opposites*. Following [4], we detect opposites by identifying pairs of events with high similarity but contrasting polarities, computed with DistilBERT for sentiment analysis[2].

[1] Specifically, the relations `xIntent`/`oWant`, and `oEffect` are clearly expressing intention and direct cause.
[2] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

## 4 Evaluation

We conducted evaluations on two widely used fact-checking datasets: AVeriTeC [12] and FEVEROUS [2]. We respectively retain only the 1,759 and 1,183 claims with a verdict in *Support, Refute, Cherry-picking* and with a causal relation.

Additionally, we constructed a manually curated **Reasoner Specific Subset (RSS)**, consisting of 86 claim-evidence pairs that contain verified use cases (e.g., causal loops or contradictions), to ensure that the mechanism should in principle be activated. We define two evaluation configurations: (1) **Tolerant**, focusing only on the system's performance when it chooses to respond (i.e. abstentions are excluded), and (2) **Strict** in which the system is expected to always produce a verdict, and abstention is treated as a false negative.

Table 2 reports the performance of our reasoning framework. Analyzing the failure cases in the RSS dataset, we observe that the Relation Extraction component exhibits notable inconsistencies. Some events are overly abbreviated, while others lack essential lexical content, making subsequent similarity and polarity computations unreliable. Moreover, the model struggles with complex linguistic structures, such as double negation. For instance, given the claim *(drinking water, intend, protect covid)* and the evidence *(hydrated, does not cause, coronavirus infection)*, the model fails to resolve the logical equivalence.

Another recurrent source of error stems from the current inability to perform type-based or ontological reasoning in the absence of explicit contextual information. For example, with the claim *(5G, causes, infertility)* against the evidence *(non-ionizing radiation, does not cause, infertility)*, the model is unable to infer that 5G, is a form of non-ionizing radiation, and shares the relevant properties.

The differences in scores between AVeriTeC and FEVEROUS may be attributed to the latter containing more straightforward claim-evidence pairs which are easier to align and reason about.

## 5 Conclusion and Future Work

By leveraging semantically refined event relationships and a structured reasoning framework, we incorporated causal reasoning into automated fact-checking, moving beyond vague representations of causality. While the proposed reasoner achieves an F1-score of approximately 50%, its main contribution is in providing structured and interpretable justifications for fact-checking verdicts while still being competitive. Rather than functioning as a standalone predictor, the system has the potential to complement existing black-box veracity classifiers.

Future work will focus on improving event extraction robustness, potentially by integrating event typing, time and space. Incorporating symbolic and ontological reasoning, fine-tuning on logical patterns, could help address the reported limitations.

| Test Set | Knowledge Source | P | R | F1-Score |
|---|---|---|---|---|
| **RSS** | LLMs | 0.55 | 0.45 | 0.50 |
| | Common Sense | 0.51 | 0.45 | 0.48 |
| **AVeriTeC (S)** | LLMs | 0.48 | 0.19 | 0.27 |
| | Common Sense | 0.54 | 0.2 | 0.29 |
| **AVeriTeC (T)** | LLMs | 0.47 | 0.35 | 0.4 |
| | Common Sense | 0.52 | 0.37 | 0.43 |
| **FEVEROUS (S)** | LLMs | 0.5 | 0.44 | 0.47 |
| | Common Sense | 0.51 | 0.44 | 0.47 |
| **FEVEROUS (T)** | LLMs | 0.52 | 0.62 | 0.56 |
| | Common Sense | 0.52 | 0.62 | 0.56 |

**Table 2: Precision, recall, and F1-score across the datasets. For RSS, the distinction between (S) = Strict abd (T) = Tolerant is unnecessary.**

## Acknowledgments

This work was supported by the European CHIST-ERA program within the ClimateSense project (Grant n° ANR-24-CHR4-0002, EPSRC EP/Z003504/1) and by the French National Research Agency (ANR) within the kFLOW project (Grant n° ANR-21-CE23-0028).

## References

[1] Marah Abdin and et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. Tech. rep. Microsoft. eprint: 2404.14219.

[2] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In *4th Workshop on Fact Extraction and VERification*. ACL, Dominican Republic, 1–13. doi:10.18653/v1/2021.fever-1.1.

[3] Zhiyun Chen, Qing Zhang, Jie Liu, Yufei Wang, Haocheng Lv, LanXuan Wang, Jianyong Duan, Mingying Xv, and Hao Wang. 2025. Counterfactual Multimodal Fact-Checking Method Based on Causal Intervention. In *Pattern Recognition and Computer Vision*. Springer Nature, Singapore, 582–595.

[4] Sebastian J. Crutch, Paul Williams, Gerard R. Ridgway, and Laura Borgenicht. 2012. The role of polarity in antonym and synonym conceptual knowledge: evidence from stroke aphasia and multidimensional ratings of abstract words. *Neuropsychologia*, 50, 11. doi:10.1016/j.neuropsychologia.2012.07.015.

[5] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10, 178–206. doi:10.1162/tacl_a_00454.

[6] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. ACL, Punta Cana, Dominican Republic, 2370–2381. doi:10.18653/v1/2021.findings-emnlp.204.

[7] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. In *28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5430–5443. doi:10.18653/v1/2020.coling-main.474.

[8] Youssra Rebboud, Pasquale Lisena, and Raphaël Troncy. 2023. Prompt-based Data Augmentation for Semantically-Precise Event Relation Classification. In *Semantic Methods for Events and Stories (SEMMES)*. CEUR, Heraklion, Greece.

[9] Youssra Rebboud, Pasquale Lisena, and Raphael Troncy. 2022. Beyond Causality: Representing Event Relations in Knowledge Graphs. In *EKAW, 23rd International Conference on Knowledge Engineering and Knowledge Management*. Springer, Bolzano, Italy. doi:10.1007/978-3-031-17105-5_9.

[10] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410.

[11] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-then Reasoning. In *33rd AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii, USA.

[12] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., New Orleans, Louisiana, USA, 65128–65167.

[13] Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal Fact Verification via Argument Structure. In *62nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1. ACL, Bangkok, Thailand, 15636–15659. doi:10.18653/v1/2024.acl-long.835.

[14] Fiona Anting Tan, Jay Desai, and Srinivasan H. Sengamedu. 2024. Enhancing Fact Verification with Causal Knowledge Graphs and Transformer-Based Retrieval for Deductive Reasoning. In *7th Fact Extraction and VERification Workshop*. ACL, Miami, Florida, USA. doi:10.18653/v1/2024.fever-1.20.