Integrating Causal Reasoning into Automated Fact-Checking

Extended Version

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

This paper is an extended version of the homonym paper accepted at ACM SAC (KNLP Track) and published in ACM proceedings.

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

1 Introduction

The rapid proliferation of both accurate and misleading content on the Web has made automated fact-checking an essential task. 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 [7]. 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 [10].

Causal reasoning has recently gained traction as a means to support factual consistency within claims [4, 19, 20]. These approaches rely on detecting cause-effect relationships between events described in the claim and those found in the evidence. However, entailment between claim and evidence is not solely based on a general notion of causality. In many real-world cases, the causal relationships between events are more nuanced and contradictions arise not from the absence of a causal chain, but from deeper incompatibilities between relational semantics. Let us illustrate with the following 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 direct match:

$$\text{Vaccine} \xrightarrow{\text{cause}} \text{No infection}$$

In this example, no such direct causal chain exists in the claim, which makes reasoning with a single causality relationship difficult. Moreover, many systems oversimplify causal relations by skipping intermediate steps. This leap from A (vaccine) to D (no infection) obscures the underlying logic and makes the reasoning path unclear 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 novel causal explanation-based verdict prediction system that relies on semantically-precise event relations – namely *cause*, *prevent*, *intend* and *enable* – derived from the FARO ontology [14], and for which a dataset has been provided in [13]. This system generates human-readable explanations based on causality extraction and a set of rules that identify the alignment or misalignment of claims and evidence. It is important to say that our approach only applies to claims that include at least one causal relation between events.

By integrating causal reasoning into the verdict prediction process, we address the limitations of existing explainability methods. The use of semantically defined relationships ensures that the explanations align with human reasoning. For this reason, our approach can also be used in combination with other fact-checking systems, with the final benefit of increasing the trust in the systems themselves. 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 and experiments are publicly available at https://github.com/ANR-kFLOW/Fact_checking_reasoner.

2 Related Work

Numerous end-to-end fact-checking systems have been developed. In [8], the authors created a system that assesses claim veracity using keyword searches for evidence and knowledge bases, relying on traditional features such as part-of-speech tags and sentiment for a 3-way classification experimenting with classifiers such as random forests and SVMs. In contrast, others employ deep neural networks for evidence selection and natural language inference, marking early examples of explainable fact-checking [18]. Popat et al. [12] use attention mechanism as a way to extract the most important words from an evidence as an explanation. The explanation gathered from the aforementioned ways is often not comprehensive [10]. Finally, other methods rely on summarization [3].

Recently, some attention has been directed towards causal reasoning in fact checking [4, 19]. In [20], the use of LLMs for causal deductive reasoning is showcased, by leveraging causal graphs and counterfactual reasoning for fact verification. These approaches are limited to what we refer to as *direct causality* within the evidence chain, without considering other types of event relations as in [14].

In [5], a method for a multi-modal detection of fake news leverages causal intervention to identify and reduce psycholinguistic bias in textual content, while counterfactual reasoning is employed to isolate and address the bias introduced by image-only cues, simulating scenarios where only visual content is available.

To the best of our knowledge, no prior research has leveraged fine-grained event-level causal relations, which are semantically richer and more nuanced than the broad or vague notions of causality typically used in explainable systems.

3 Reasoning Rules

In this section, we introduce relation-based reasoning rules to deduce the most probable verdict for a given claim, knowing the evidence. These rules are intended to be found: (1) between events in the claim; (2) between events in each evidence; (3) between one event in the claim and one event in the evidence (or vice-versa).

3.1 Types of Relationships

The FARO ontology [14] defines a set of semantically precise event relations. This ontology not only provides a textual definition to several event relations, but already defines logical axioms such as transitivity or disjunction using the OWL representation language.

We focus our work on the four FARO relations, covering a broad range of cases:

- **direct-cause**: a relationship between an event and its effect.
  Ex: *the earthquake has left behind dozens of deaths in Japan.* (earthquake, causes, deaths)
  Transitive[1] | **Disjointwith**: does-not-cause.
- **prevents**: connect an event instance with the event for which is the cause of not happening.
  Ex: *Boy Scouts of America are committed to act against racial injustice.* (act, prevents, racial injustice)
  SuperProperty: does-not-cause | **Disjointwith**: cause
- **intends-to-cause**: establish a link between an event and its intended effect, regardless of whether the desired outcome is ultimately achieved.
  Ex: *The company implemented a marketing strategy to boost the product sales.* (implementing, intends to cause, boost)
  Disjointwith: does-not-cause, prevent.
- **enables**: connect a condition with the event it is contributing to perform as an enabling factor.
  Ex: *having access to reliable internet grants students access to online courses.* (access to reliable internet, enables, access to online courses)
  Disjointwith: does-not-cause, prevent.

3.2 Rules

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 "$A \longrightarrow B$" in the claim and C and D with the relationship "$C \longrightarrow D$" in the evidence. In the following subsections, we outline four key scenarios, all illustrated with examples.

3.2.1 Logical alignment. The *logical alignment* scenario is verified if the claim and the evidence include the same (or similar, or transitively-linked) events, which 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:

- C is similar to A and D is similar to B;
- 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.

Claim: *Sumo wrestler Toyozakura Toshiaki committed match-fixing, ending his career in 2011.*

Evidence: *He was forced to retire in April 2011 after an investigation by the Japan Sumo Association found him guilty of match-fixing.*

$$\begin{array}{llll} A: & \textit{Committed match fixing} & C: & \textit{Investigation} \\ B: & \textit{Ending} & D: & \textit{Retire} \end{array}$$

$$\begin{array}{lll} (1) & A \xrightarrow{\text{causes}} B & (\textit{Direct cause from the claim}) \\ (2) & C \xrightarrow{\text{causes}} D & (\textit{Direct cause from the evidence}) \\ (3) & A \xrightarrow{\text{causes}} C & (\textit{Direct cause linking claim and evidence}) \\ (4) & B = D & (\textit{Equivalence from the evidence}) \end{array}$$

Deduction:

$$1.\quad A \xrightarrow{\text{causes}} C \text{ and } C \xrightarrow{\text{causes}} D \implies A \xrightarrow{\text{causes}} D \quad \text{(by transitivity)}$$

$$2.\quad B = D \implies A \xrightarrow{\text{causes}} B$$

Conclusion: $A \xrightarrow{\text{causes}} B$ is confirmed through transitivity (via evidence). The evidence and the claim are logically consistent, demonstrating proper alignment. Figure 1 illustrates this scenario.

[1] We adopt the commonly held assumption that direct causality is transitive. Despite ongoing debates [11], we align with the prevailing perspective in the literature that supports this assumption.

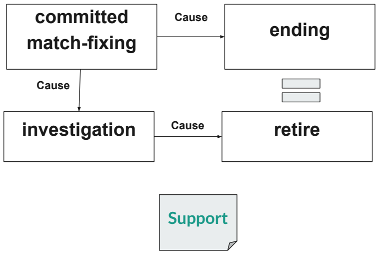

Figure 1: An example of a logical alignment.

3.2.2 *Logical Misalignment.*

In the *logical misalignment* scenario, 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.

Claim: *Exercising daily causes muscle fatigue over time.*

Evidence: *Research shows that daily low-intensity exercise activates recovery mechanisms in the body, preventing the onset of chronic muscle fatigue and improving overall stamina instead.*

$$\begin{array}{llll} A: & \textit{Exercising daily} & C: & \textit{Activates recovery mechanisms} \\ B: & \textit{Muscle fatigue} & D: & \textit{Prevents muscle fatigue} \end{array}$$

$$\begin{array}{ll} (1)\ A \xrightarrow{\text{causes}} B & (\textit{from the claim}) \\ (2)\ A \xrightarrow{\text{causes}} C & (\textit{from the evidence}) \\ (3)\ C \xrightarrow{\text{prevents}} D & (\textit{from the evidence}) \\ (4)\ B = D & \end{array}$$

Deduction: Using the above relationships:

1. $A \xrightarrow{\text{causes}} C$

2. $C \xrightarrow{\text{prevents}} D \implies C \xrightarrow{\text{causes}} \neg D$

   $\implies A \xrightarrow{\text{causes}} \neg D \implies A \xrightarrow{\text{prevents}} D$

3. $A \xrightarrow{\text{causes}} B$ and simultaneously $A \xrightarrow{\text{prevents}} B$ (Contradiction).

Contradiction: A single event (A: *Exercising daily*) cannot simultaneously *cause* and *prevent* the same outcome (B: *Muscle fatigue*).

Conclusion: The claim and evidence are logically inconsistent, as the evidence contradicts the claim's assertion (Figure 2).

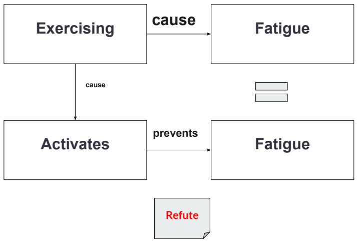

Figure 2: An example of a logical misalignment.

3.2.3 *Causal loops.*

We check for a *closed causal loop* 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.

Claim: *Poor infrastructure causes economic decline.*

Evidence: *transportation inefficiencies leads to supply chain disruptions and reduced economic activity.*

$$\begin{array}{llll} A: & \textit{Poor infrastructure} & C: & \textit{Transportation inefficiencies} \\ B: & \textit{Economic decline} & D: & \textit{Supply chain disruptions} \end{array}$$

$$\begin{array}{ll} (1)\ A \xrightarrow{\text{causes}} B & (\textit{from the claim}) \\ (2)\ A \xrightarrow{\text{causes}} C & (\textit{from the evidence}) \\ (3)\ C \xrightarrow{\text{causes}} D & (\textit{from the evidence}) \\ (4)\ D \xrightarrow{\text{causes}} B & (\textit{from the evidence}) \end{array}$$

Deduction:

1. $A \xrightarrow{\text{causes}} C$ and $C \xrightarrow{\text{causes}} D \implies A \xrightarrow{\text{causes}} D$
2. $A \xrightarrow{\text{causes}} D$ and $D \xrightarrow{\text{causes}} B \implies A \xrightarrow{\text{causes}} B.$

Conclusion: The claim $A \xrightarrow{\text{causes}} B$ is supported by the evidence through a causal loop.

3.2.4 *Cherry-picking.*

The practice commonly addressed as *cherry-picking* consists of 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 $\langle \text{sub}, \text{rel}, \text{obj} \rangle$ 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. sub_1 vs. sub_2, obj_1 vs. 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.

Evidence 1: *Frequent testing of the entire population would help identify so-called hidden carriers, individuals infected with SARS-CoV-2 [...] Identifying these silent spreaders could help public health workers be more effective at contract tracing by identifying others who have been exposed and may require quarantine.*

Evidence 2: *Testing the entire population would undoubtedly identify a large number of such individuals, unnecessarily sidelining them from work and society.*

While the event *Testing* from the first evidence is equivalent to *Testing* from the second evidence, the second components of the relation differ significantly. Not only are they dissimilar, but they also have opposite polarities, the first being positive (tracing), while

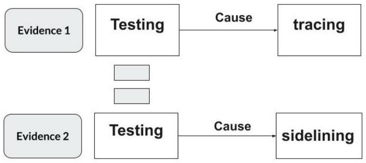

Figure 3: An example of a Cherry-Picking Scenarios.

the second being negative (sidelining). This discrepancy may raise concerns about a potential cherry-picking scenario (Figure 3).

4 Methodology

Our pipeline (Figure 4) begins with event relation extraction conducted separately within the claim and evidence (Section 4.1). We describe the approaches used to extract fine-grained causal relationships across claims and evidence (Section 4.2) and to distinguish between similar, dissimilar, and opposite events (Section 4.3). These modules are then combined in our reasoning approach (Section 4.4).

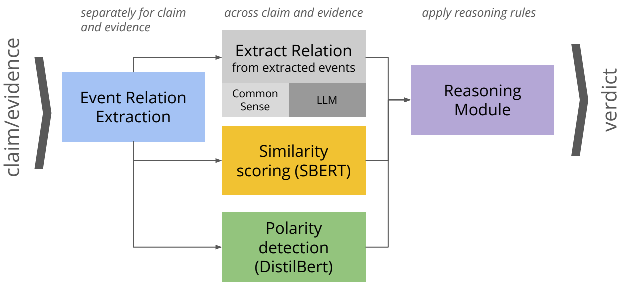
separately for claim and evidence
across claim and evidence
apply reasoning rules
claim/evidence
Event Relation Extraction
Extract Relation from extracted events
Common Sense
LLM
Similarity scoring (SBERT)
Polarity detection (DistilBert)
Reasoning Module
verdict

Figure 4: Overview of our proposed pipeline

4.1 Causality Extraction within Claim/Evidence

Causality extraction is initially performed within the same context (either the claim or the evidence) using an automatic filtering process followed by manual validation. The system leverages the sequence-to-sequence model REBEL [9], an auto-regressive architecture comprising an encoder and a decoder layer. We trained REBEL on a previously available annotated news dataset [13] that includes 2,696 training, 265 development, and 461 testing sentences annotated with events and the four studied fine-grained causal relations. It processes raw text to produce the corresponding triplet *(Subject, Relation, Object)*. This system obtained a combined F1-score of 0.82 (0.89 for identifying the subject, 0.75 for the object).

4.2 Causality Extraction across Claim/Evidence

In Figure 5, we aim to identify the prevention relationship between the event *"limit all non essential interactions"* in the evidence and the event *"death"* in the claim. We have experimented with two different strategies: one relying on common sense knowledge bases and one relying on Large Language Models (LLM).

4.2.1 Common Sense-based Causality Extraction. The Atlas of Machine Commonsense (ATOMIC) [16] is a large knowledge graph that contains over 877k inferential tuples representing common-sense situations, related by known relations. We identified an overlap between its relations and causality relations. Specifically, the relations `xIntent`/`oWant`, and `oEffect` are clearly expressing intention and direct cause. We isolated the ATOMIC tuples involving these properties, creating a reference dataset.

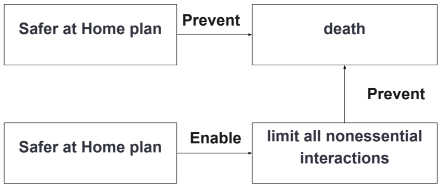

Figure 5: An example of refined causality extraction between events across the Claim and the Evidence

To address the absence of *enabling* and *prevention*, we have expanded the dataset by generating new examples using a LLM. Specifically, we used Zephyr, a fine-tuned version of Mistral 7B. As manual assessment, we randomly selected and judge 100 examples from each relation type (*prevention*, *enable*), revealing that a significant portion of the generated examples (around 97% for prevent and 76% for enable) were accurate enough to train our model.

To handle the cases where there is no relation between the two events, we use negative sampling. We select 50% of the whole dataset while stratifying the relation type, and we organize the dataset as a set of pairs of rows, so that a pair is two rows or two triples. Then, we swap randomly either the subject of the pair or the object. 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

4.2.2 LLM-based Causality Extraction. We experimented with the *Phi-3-Medium-4K-Instruct* model, which demonstrates good performance in common-sense reasoning, while at the same time requiring fewer parameters and less computational effort [1]. The following prompt was employed in a few-shot setting:

Event Relation Extraction Prompt

`User:` Knowing that I want to extract refined causal relations between two given events, and it only can be *cause*, *intend*, *prevent*, *enable*, or *no relation*, you have to answer only with the relation name, no explanation. What will be the relation between earthquake and death?
`Assistant:` cause.
`User:` What about relation between {event1} and {event2}?

To evaluate the performance of the LLM in extracting fine-grained causality between events across claims and evidence, we extracted and manually assessed 40 samples from both claims and evidences. Out of the 40 samples, 33 were correctly processed (82.5%).

4.3 Similarity, Dissimilarity, and Opposites

To determine if two events are the same, we rely on sentence similarity and dissimilarity, computed using SentenceBERT [15].

We evaluate two configurations: (1) *Events only* and (2) *Events with context*, that is the concatenation of event spans and their original claim/evidence sentence. For cases where the events are an exact match (same surface form), we rely on the *Events only* configuration, and otherwise apply *Events with context.* To illustrate this approach, we use of the following examples.

(1) **Claim:** *Dr. Qadir went on hunger strike when in prison [...]; then he was released from custody on January 25, 2006, as a result of **efforts by special envoy of the Austrian foreign ministry**, Gudrun Harrer, a journalist.*
Evidence: *He was released from custody on January 25, 2006, as a result of **efforts** by special envoy of the Austrian foreign ministry, Gudrun Harrer.*
(2) **Claim:** *Sumo wrestler Toyozakura Toshiaki committed match-fixing, **ending** his career in 2011.*
Evidence: *He was forced to **retire** in April 2011 after an investigation by the Japan Sumo Association found him guilty of match-fixing.*
(3) **Claim:** *The **drought** caused severe crop failures.*
Evidence: *Because of the prolonged dry **conditions**, agricultural yields were dramatically lower than usual.*

In these examples, we have an alignment between events in the claim and events in the evidence. For each case, the *Events with context* configuration produced the higher similarity score, as shown in Table 2. Based on empirical results on the entire dataset, we set the threshold for event similarity to **0.54**, as events with similarity above this value are considered 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.* According to [6], antonyms can be as similar or even more similar than synonyms, aside from their polarity (Pol). This insight suggests that we can detect opposites by identifying pairs of events with high similarity but contrasting polarities.

Input	Ex1	Ex2	Ex3
Events only	0.3235	0.2448	0.2589
Events + context	**0.7632**	**0.6709**	**0.6533**

Table 2: Cosine similarity for the three examples described above.

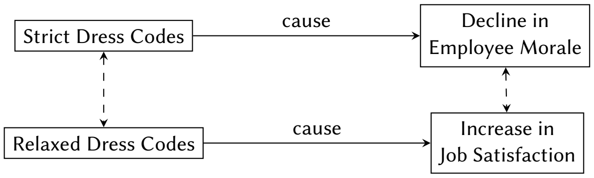

Figure 6: Two pairs of "opposite" triples.

To investigate this, we sampled five pairs of claims and corresponding evidence in which the statements contradict each other, and the events involved are opposites. We computed similarity and polarity under three configurations:

- Evaluating events in isolation,
- Evaluating whole claim text vs whole evidence text, and
- Evaluating the full triple text *(sub,rel,obj).*

We discuss the following example, represented in Figure 6:
Claim: *A study released on November 15, 2023, found that companies with strict dress codes experience a decline in employee morale.*
Evidence: *The Workplace Institute surveyed 150 firms with relaxed dress codes in October 2023 and found that employees reported a 25% higher job satisfaction rate than those at 100 companies enforcing formal attire.*

We determined which configuration best captured the correct polarity alongside high similarity. The configuration that produced the correct opposing polarities with strong similarity scores was ultimately chosen to identify and confirm opposite relationships.

Similarity is computed as the cosine similarity between the embeddings of two input events obtained from SentenceBERT. For polarity detection, we employed the DistilBERT base uncased model fine-tuned for sentiment analysis[2]. The model takes the input text and outputs a polarity classification: N (Negative) or P (Positive).

Evaluating the complete triple yields the highest percentage of correct polarities, and the second highest similarity scores making it the most effective among the considered configurations (Table 3).

Based on the empirically determined similarity threshold θ and the opposites check, which suggests that opposites are semantically similar but exhibit different polarities, we define the following rules. Given two text inputs T_1 and T_2:

- **Is-Similar**$(T_1, T_2) \implies$
Similarity$(T_1, T_2) > \theta$ **and** Pol$(T_1) = $ Pol(T_2) (PP or NN).
- **Is-Dissimilar**$(T_1, T_2) \implies$ Similarity$(T_1, T_2) < \theta$.
- **Opposites**$(T_1, T_2) \implies$
Similarity$(T_1, T_2) > \theta$ **and** Pol$(T_1) \neq$ Pol(T_2) (PN or NP)

4.4 Reasoning Approach

The reasoner analyzes the claim and evidence by checking the relationships between their events. It performs the following tasks:

[2] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

Comparison	Correct Polarity	Average Similarity
Claim vs Evidence	20%	**0.76**
Claim Event vs Evidence Event	50%	0.50
Triple Claim vs Triple Evidence	**80%**	0.61

Table 3: Comparison of correct polarity and average similarity across different levels of analysis.

(1) **Causal Loop check** (Figure 7) verifies if the events form a closed causal cycle, indicating support for the claim.
(2) **Similarity and Relationship check** (Figure 8) compares the relationships and similarities between events to determine alignment or contradiction.
(3) **Cherry-picking check** (Figure 9) identifies inconsistencies or selective usage of evidence that may bias the verdict.

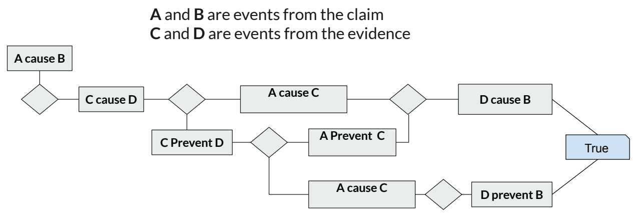
A and **B** are events from the claim
C and **D** are events from the evidence
A cause B
C cause D
A cause C
D cause B
C Prevent D
A Prevent C
True
A cause C
D prevent B

Figure 7: Implementation structure of the Causal Loop check.

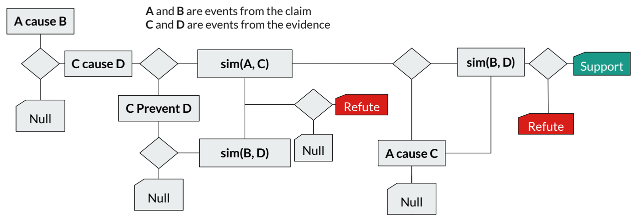

Figure 8: Implementation structure of a Similarity and Relationship check. *Is-similar* is shortened to *sim* for readability

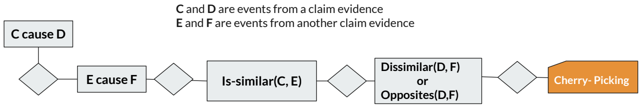
C and D are events from a claim evidence
E and F are events from another claim evidence
C cause D
E cause F
Is-similar(C, E)
Dissimilar(D, F)
or
Opposites(D,F)
Cherry- Picking

Figure 9: Implementation structure of a cherry picking scenario check.

The assigned labels are categorized as in AVeriTeC (see Section 5.1):

- **Supported:** When the evidence fully supports the claim.
- **Refuted:** When the evidence contradicts the claim.
- **Conflicting Evidence/Cherrypicking:** When the evidence neither fully supports nor completely refutes the claim, but exhibits conflicting information.

5 Evaluation

This section outlines the evaluation datasets and the strategies used to assess our reasoning framework. We explain the filtering criteria applied to each dataset and the evaluation setups adopted for performance analysis.

5.1 Evaluation Datasets

To assess the effectiveness of our reasoning framework, we conducted evaluations on two widely used fact-checking datasets: AVeriTeC [17] and FEVEROUS [2]. Additionally, we constructed a manually curated subset of relevant use cases.

5.1.1 AVeriTeC. It consists of 4,568 real-world claims, each paired with question–answer evidence and textual justifications used to determine verdicts, and annotated with one of four verdict labels:

- **Supported:** The evidence fully supports the claim.
- **Refuted:** The evidence directly contradicts the claim.
- **Conflicting Evidence (Cherry-picking):** The evidence presents conflicting information neither fully supporting nor fully refuting the claim.
- **Not Enough Evidence**: The evidence is insufficient to make a conclusive judgment about the claim veracity.

In our evaluation, we use the training subset of AVeriTeC, but we retained only claims linked to informative textual answers, so excluding boolean and unanswerable cases. We also exclude the *Not Enough Evidence* portion of the dataset, since our reasoning system is not designed to produce this type of verdict. This leaves 2998 claims.

5.1.2 FEVEROUS dataset. It contains 87,026 claims annotated with evidence sourced from Wikipedia. Each claim is labelled as *supports*, *refutes*, or *not enough info*. Evidence may include textual sentences or table cells, along with annotator metadata – e.g., query actions, page clicks, evidence types.

We randomly selected a subset of 4,392 claims for our experiments, to have a dataset similar to the AVeriTeC in volume. We retained only claims supported by fully textual evidence and excluded those referencing table cells, yielding a filtered subset suitable for text-only based reasoning. We also excluded the *Not enough info* part since it is not handled by the reasoner. We then filtered out claims without causal relations, retaining a total of 1,183 claims for evaluation (705 *supports* and 478 *refutes*).

5.1.3 Reasoner Specific Subset (RSS). Our reasoning framework is based on a set of rule-based mechanisms operating over event relations, similarity, and polarity. While this approach allows for explicit and interpretable inference, it does not guarantee coverage of all examples within the datasets, as not all claim-evidence pairs contain use cases compatible with the system's reasoning rules.

As additional evaluation, we constructed a controlled subset consisting of claim-evidence pairs that contain verified use cases: while we do not assume that the reasoner will always respond correctly (the final outcome also depends on other components, e.g.

similarity scoring and polarity detection), we ensure that a valid use case is present and the mechanism should in principle be activated.

We randomly sampled 765 claims from the AVeriTeC dataset. For each claim-evidence pair, we checked whether it contained a valid use case for reasoning, e.g., causal loops or contradictions. We only include in this dataset the 86 pairs (across 60 unique claims) presenting a valid use case.

Table 4 shows the stats for each dataset after filtering.

Filtering Step	AVeriTeC	FEVEROUS
Total unique claims	2998	1736
Total answers / total evidences	8479	3836
Answer Type Distribution (AVeriTeC)		
Extractive	4571	–
Abstractive	2225	–
Boolean	1297	–
Unanswerable	386	–
Claims *excluding Boolean and Unanswerable*	2783	1736
Claims with no relation	850	44
Claims *excluding "not enough evidence"*	1759	1183
Label Distribution		
Refuted / REFUTES	1066	478
Supported / SUPPORTS	581	705
Conflicting Evidence / Cherrypicking	112	–

Table 4: Filtering steps and label distributions for the datasets before running the reasoning pipeline

5.2 Evaluation Strategy

In our evaluation, we define two distinct configurations for computing performance metrics:

- **Configuration 1 – Tolerant:** It adopts a lenient evaluation strategy, focusing only on the system's performance when it chooses to respond.
  - **Recall** is defined as the proportion of cases where the reasoner successfully produces a verdict (i.e., does not abstain), relative to the total number of evaluated cases.
  - **Precision** measures the proportion of correct verdicts among those that were actually produced (i.e., abstentions are excluded).
  - **Abstentions** (*None* outputs) are excluded from metric computation. This is because they may arise from either genuinely irrelevant inputs or from missed reasoning opportunities due to limitations in similarity matching or claim–evidence alignment.
- **Configuration 2 – Strict:** This configuration uses a more rigorous evaluation policy. It is primarily applied to the manually verified RSS, where each item has been checked to ensure that it is reasoning-relevant.
  - The system is expected to always produce a verdict. Every abstention (*None*) is treated as a **false negative (FN)**, thus reducing recall.
  - **Recall** is computed as the number of correct verdicts over the total number of evaluated cases, including abstentions.

In AVeriTeC and FEVEROUS, all examples—including abstentions—are evaluated, with abstentions counted as false negatives. This strict setup tests the model's robustness in ambiguous cases.

5.3 Results and Discussion

Table 5 reports the performance of our reasoning framework across different configurations, datasets, and knowledge sources for causality extraction across claims and evidences. On the reasoner specific subset (RSS), the system achieves an F1-score of 0.50 with LLMs and 0.48 with common-sense knowledge bases ERE. In 50% of the cases the model either does not provide an answer or the answer is wrong.

Upon analyzing the failure cases in the RSS dataset, we observe that the Event Relation Extraction component exhibits notable inconsistencies. Some events are overly abbreviated, while others lack essential lexical content, making subsequent similarity and polarity computations highly 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 due to nested negation and lexical variation.

Another recurrent source of error stems from the reasoner's current inability to perform type-based or ontological reasoning. 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. In the absence of explicit contextual or ontological information, the system fails to capture the semantic similarity required for correct inference.

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 5: Precision, recall, and F1-score for each knowledge source across the different evaluation datasets. RSS refers to the Reasoner-Specific Subset, in which tolerant evaluation was unnecessary as all examples are guaranteed to trigger reasoning. (S) stands for the Strict evaluation while (T) refers to the Tolerant one.

The evaluation of AVeriTeC and FEVEROUS reveals notable differences in performance between strict and tolerant settings, offering distinct insights into how each dataset responds to these configurations. For AVeriTeC, the strict setting leads to a sharp decline in recall for both sources, a result that aligns with expectations since abstentions—classified as "None" are treated as false negatives. Conversely, the tolerant configuration paints a more favorable picture: when the system does provide a response, its predictions tend to be accurate, as reflected in the higher F1-scores. In the case of FEVEROUS, performance remains more consistent across both reasoning approaches. Under the strict setting, both reasoners achieve an F1-score of 0.47, while the tolerant setting sees this figure rise to 0.56 for both, with recall reaching as high as 0.62. This pattern indicates that FEVEROUS likely contains more straightforward, fact-based claim-evidence pairs, easier to align and reason about, resulting in more stable performance across configurations.

Comparison with GPT-4o mini. To better assess our work, we compared our rule-based reasoner against a powerful large language model (LLM), GPT-4o mini, for the RSS dataset. GPT-4o mini was provided with each claim together with the complete set of corresponding evidences as input. The model achieved a macro F1-score of 0.67 across the three labels, outperforming our reasoner, which reached a score of 0.50. This corresponds to an improvement of approximately +17%. However, GPT-4o mini may have been previously exposed to similar samples from the RSS dataset during training, a contamination that could lead to an overestimation of its real performance. Despite its advanced reasoning abilities, the model still struggles to achieve perfect consistency, confirming the intrinsic complexity of causal relation reasoning in factual verification.

6 Conclusion and Future Work

This study presents a novel approach for incorporating causal reasoning into automated fact-checking. By leveraging semantically refined event relationships and a structured reasoning framework, our system addresses the prevalent limitation of causal interpretability in existing methods—moving beyond vague or shallow representations of causality.

While the proposed reasoner achieves an F1-score of approximately 50%, its main contribution is not in outperforming existing models in terms of raw accuracy, but in providing structured and interpretable justifications for fact-checking verdicts while still being very competitive. Rather than functioning as a standalone predictor, the system has the potential to complement existing black-box veracity classifiers, including those offering limited explainability based on vague or underspecified causal links. By surfacing explicit causal links, polarity mismatches, and logical inconsistencies, it provides valuable explanatory signals to support or challenge automated decisions.

Our error analysis highlights several avenues for improvement as future work. First, inconsistencies in ERE outputs, such as incomplete or overly abstract event representations significantly hinder downstream similarity and polarity computations. Future work will focus on improving event extraction robustness, potentially by integrating event typing. The event representation should also aggregate the information about time and space.

The current system struggles with complex linguistic phenomena such as double negation and implicit entailment. Incorporating symbolic reasoning layers or transformer-based inference modules fine-tuned on logical patterns could help address this limitation.

Finally, the system lacks the capacity for ontological reasoning, which is crucial for handling type-based mismatches (e.g., recognizing that 5G is a subclass of non-ionizing radiation). Future directions include enriching the model with external ontologies or commonsense knowledge bases to support type inference and contextual disambiguation. Such enhancements would improve both the coverage and reliability of the causal reasoning pipeline in real-world fact-checking scenarios.

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 4^{th} *Workshop on Fact Extraction and VERification.* ACL, Dominican Republic, 1–13. doi:10.18653/v1/2021.fever-1.1.

[3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *58th Annual Meeting of the Association for Computational Linguistics.* ACL, Online, 7352–7364.

[4] 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.

[5] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In 61^{st} *Annual Meeting of the Association for Computational Linguistics.* Vol. 1. ACL, Toronto, Canada, 627–638. doi:10.18653/v1/2023.acl-long.37.

[6] 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.

[7] 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.

[8] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '17). Association for Computing Machinery, Halifax, NS, Canada, 1803–1812. isbn: 9781450348874. doi:10.1145/3097983.3098131.

[9] 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.

[10] 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.

[11] Neil McDonnell. 2018. Transitivity and proportionality in causation. *Synthese*, 195, 3, 1211–1229. doi:10.1007/s11229-016-1263-1.

[12] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *2018 Conference on Empirical Methods in Natural Language Processing.* ACL, Brussels, Belgium, 22–32. doi:10.18653/v1/D18-1003.

[13] 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.

[14] 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.

[15] 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.

[16] 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 *33^{rd} AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii, USA.

[17] 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.

[18] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *25^{th} ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (KDD '19). ACM, Anchorage, AK, USA, 395–405. doi:10.1145/3292500.3330935.

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

[20] 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 *7^{th} Fact Extraction and VERification Workshop*. ACL, Miami, Florida, USA. doi:10.18653/v1/2024.fever-1.20.