DEDUCE: ROBUST FACT-CHECKING UNDER KNOWLEDGE POISONING VIA CLAIM DECOMPOSITION AND EVIDENCE REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated fact-checking systems traditionally operate under the "clean-knowledge" assumption: the premise that retrieved evidence is inherently reliable. However, in open-domain environments, knowledge bases may contain fabricated or misleading content, causing substantial degradation when misinformation is retrieved as evidence. We formalize the task of Fact-Checking under Contamination (FCC) and propose DEDUCE, a robust verification framework designed for contaminated knowledge environments. DEDUCE combines claim decomposition, claim-conditioned stance aggregation, and dynamic passage-level reliability estimation in a closed-loop retrieval process, where learned reliability signals are fed back into subsequent retrieval through reranking. To rigorously evaluate this setting, we introduce a controlled stress test with topically aligned, adversarially generated distractors and complement it with naturally occurring misinformation. Across seven benchmarks and four open-source LLM backbones, contamination substantially degrades standard fact-checking pipelines, while DEDUCE consistently recovers a significant portion of the lost performance. Additional comparisons on HoVer and EX-FEVER show that DEDUCE also outperforms purpose-built poisoning defenses, while further stress tests reveal that coordinated and reputation-based misinformation remain challenging when misleading evidence becomes highly consistent or inherits previously accumulated reliability. These results demonstrate the importance of explicitly modeling evidence reliability and retrieval feedback for robust fact checking under contaminated knowledge.

1 INTRODUCTION

The proliferation of misinformation poses a growing challenge to reliable information access. Automated fact-checking systems increasingly combine retrieval mechanisms (Lewis et al., 2020; Thorne et al., 2018) with Large Language Models (LLMs) to assess claim veracity, but they often assume that the underlying knowledge base (KB) is reliable. In practice, retrieved evidence may contain accurate, misleading, and fabricated documents, and such contamination can substantially degrade verification performance.

In this work, we formalize Fact-Checking under Contamination (FCC), where fact-checking operates over an unreliable KB. We study this setting using two complementary sources of misinformation. First, we construct controlled contaminated KBs by generating topically aligned adversarial passages from gold evidence and false claims, and independently validate the generated misinformation. Second, we evaluate on RAGuard, which contains naturally occurring misinformation from political and social media sources. Together, these settings capture both targeted adversarial contamination and naturally occurring noisy evidence.

To address FCC, we propose DEcomposition and Dynamic Updating for Contaminated Evidence (DEDUCE). DEDUCE first decomposes complex claims into simpler subclaims and retrieves evidence separately for each, reducing dependence on any single noisy retrieval result. It then uses an LLM verifier to classify retrieved passages as supporting, refuting, or unrelated. Based on these claim-conditioned stance observations, DEDUCE maintains passage-level reliability estimates and feeds them back into subsequent retrieval through reranking. Thus, verification outcomes influence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

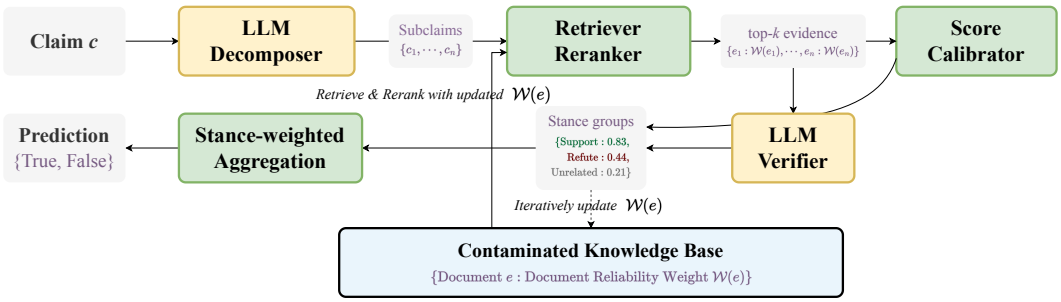


Figure 1: Overview of DEDUCE. A claim is first decomposed into simpler subclaims. For each subclaim, the retriever obtains the top-k relevant documents from a contaminated KB. The LLM verifier classifies each document as supporting, refuting, or unrelated to the subclaim. The resulting stance groups are used for subclaim verification and for updating document reliability weights, which are then incorporated into evidence reranking for subsequent predictions.

future evidence selection, forming a query-dependent retrieval-feedback loop rather than a static reliability aggregation procedure.

We evaluate DEDUCE with four open-source LLMs across seven datasets. Contamination substantially degrades standard fact-checking pipelines, while DEDUCE consistently recovers a considerable portion of the lost performance. On HoVer and EX-FEVER, DEDUCE also outperforms adapted classical truth-discovery methods and purpose-built robust-RAG defenses, while combining DEDUCE with a filtering-based defense yields further gains. Component ablations show that claim decomposition, stance grouping, and reliability updating each contribute to performance, and additional experiments demonstrate the importance of cross-claim reliability accumulation and retrieval feedback. Stress tests further reveal clear limitations under coordinated misinformation, mixed-truth passages, and adaptive reputation poisoning, where textual stance consistency alone may be insufficient.

Our contributions are threefold:

1. **FCC formulation and evaluation.** We formalize fact-checking over contaminated KBs and evaluate robustness using both controlled adversarial misinformation and naturally occurring misleading evidence.
2. **Contamination-aware fact checking with DEDUCE.** We introduce a framework combining claim decomposition, claim-conditioned stance aggregation, passage-level reliability estimation, and retrieval-feedback reranking.
3. **Comprehensive robustness analysis.** Across seven datasets and four LLM backbones, we compare against recent fact-checking systems, classical truth-discovery methods, and robust-RAG defenses, while analyzing component contributions and challenging failure modes.

2 Related Work

Research on reliable fact-checking spans several interconnected areas, including automated fact-checking, misinformation and adversarial evidence, robust retrieval, and truth discovery. Fact-checking examines how to assess the veracity of a claim given external evidence, while research on misinformation and retrieval robustness studies how misleading or corrupted evidence can affect downstream reasoning. Our work lies at the intersection of these directions, explicitly addressing FCC, where the retrieved evidence itself may contain adversarial misinformation.

Automated Fact-Checking. Automated fact-checking has been a longstanding focus within natural language processing. Benchmarks such as FEVER (Thorne et al., 2018), HoVer (Jiang et al., 2020), and EX-FEVER (Ma et al., 2023) provide large-scale collections of labeled claims with associated evidence, facilitating systematic evaluation of retrieval and reasoning components. Recent work increasingly incorporates LLMs to handle complex claims requiring multi-step reasoning. GraphCheck (Jeon & Lee, 2025) represents claims as entity-relation graphs and decomposes them into simpler subclaims for structured verification, while ProgramFC (Pan et al., 2023) decomposes complex claims into executable reasoning programs. PACAR (Zhao et al., 2024) strengthens fact-checking through planning and customized action reasoning, and LoCal (Ma et al., 2025) employs

logical and causal reasoning with multiple LLM agents. These methods primarily improve the reasoning process under retrieved evidence. In contrast, DEDUCE focuses on the complementary problem of maintaining reliable verification when the retrieved evidence corpus itself contains conflicting or adversarial passages. Accordingly, we additionally evaluate PACAR and LoCal under both clean and contaminated retrieval in our experiments.

Misinformation and Adversarial Retrieval. A growing body of work studies how misleading evidence affects retrieval-augmented models. Weller et al. (2022) show that injecting disinformation into open-domain QA corpora can substantially degrade performance and propose a redundancy-based defense that searches for consistent answers across retrieved contexts. Their approach increases evidence redundancy through query rewriting, whereas DEDUCE maintains persistent passage-level reliability signals and feeds them back into subsequent retrieval. PoisonedRAG (Zou et al., 2025) further demonstrates that manipulating the retrieval corpus can directly compromise downstream generation. Related work also examines robustness to misleading context in specialized domains, including MedMisBench (Zhou et al., 2026) and targeted misinformation attacks on medical LLMs (Han et al., 2024). RAGuard (Zeng et al., 2025) provides a complementary fact-checking setting containing naturally occurring misleading evidence.

Recent work has also proposed defenses specifically for unreliable retrieval. RobustRAG (Xiang et al., 2024) uses an isolate-then-aggregate strategy to reduce the influence of corrupted content, TrustRAG (Zhou et al., 2025) combines embedding-based clustering with consistency-driven conflict resolution, and Astute RAG (Wang et al., 2025) addresses retrieval conflicts by integrating retrieved evidence with parametric knowledge. We adapt these methods to our binary fact-checking setting under the same contaminated corpora and retrieval budget, and additionally evaluate a combined filtering-and-reweighting variant to study the complementarity between filtering-based defenses and the reliability-aware retrieval feedback in DEDUCE.

Truth Discovery and Reliability Estimation. The document reweighting mechanism in DEDUCE is closely related to classical truth discovery, which jointly estimates source reliability and the truthfulness of conflicting observations. TruthFinder (Yin et al., 2007), Investment (Pasternack & Roth, 2010), and AccuSim (Dong et al., 2009) provide representative approaches for propagating confidence between sources and claims, establishing the general principle that reliability estimation and truth inference can reinforce one another.

DEDUCE builds on this principle rather than treating reliability updating itself as novel. However, its setting differs from classical truth discovery in several ways: evidence is retrieved dynamically for each subclaim, reliability is maintained at the passage level, and observations come from claim-conditioned LLM judgments of support, refutation, or irrelevance. Most importantly, the resulting reliability estimates are fed back into subsequent retrieval through reranking, allowing verification outcomes to influence future evidence selection. We therefore position DEDUCE as integrating passage-level reliability estimation into a query-dependent retrieval-feedback loop for FCC. In our experiments, we adapt TruthFinder, Investment, and AccuSim to the same retrieved candidates and stance observations to isolate the effects of reliability aggregation and retrieval feedback.

3 METHODS

We propose DEDUCE, a stance-based verification framework that combines claim decomposition with dynamic passage-level reliability estimation and retrieval-feedback reranking. Figure 1 provides an overview of DEDUCE. Section 3.1 formally defines the problem setting. Section 3.2 presents our procedure for constructing controlled misinformation in the KB. Section 3.3 describes the claim decomposition strategy. Section 3.4 introduces stance grouping, document reliability updating, and reliability-aware evidence reranking for robust verification.

3.1 PROBLEM FORMULATION

Given a set of claims C and a contaminated KB $\tilde{K}$, DEDUCE integrates a large language model V as a verifier and a search engine R as a retriever. The goal is to predict a binary label $Y \in \{\text{TRUE}, \text{FALSE}\}$ for each claim based on evidence retrieved from $\tilde{K}$.

We consider a setting where the KB contains both reliable and misleading documents. This differs from standard fact-checking benchmarks, where the knowledge source is usually assumed to be clean. DEDUCE is designed for cases where truthful evidence remains retrievable, but the retrieved evidence pool may also contain topically aligned misinformation. This assumption is reasonable for large corpora such as Wikipedia, where factual information is often repeated across multiple documents. At the same time, we do not assume that every retrieved document is reliable, nor that misinformation can be identified from surface form alone. Instead, DEDUCE uses aggregate stance consistency across retrieved evidence to estimate both subclaim veracity and passage-level reliability.

For a subclaim c, the retriever returns a set of evidence documents E with retrieval scores $\rho(e)$. The verifier then partitions these documents into three stance groups: supporting documents S, refuting documents F, and unrelated documents U. Supporting and refuting documents are used to determine the veracity of the subclaim and to update document reliability statistics. Unrelated documents are excluded from stance aggregation and reliability updating.

3.2 Misinformation KB Construction

Designing effective misinformation is challenging. Naive strategies such as direct negation or entity replacement often produce distractors that are either too obvious, insufficiently diverse, or not strictly false. For example, replacing ***painter*** with ***engineer*** in *"Leonardo da Vinci is a painter"* does not necessarily make the resulting statement false, since da Vinci was both. Our experiments also show that such strategies only weakly degrade fact-checking performance under contaminated KBs.

Algorithm 1 Subclaim Verification

Input: contaminated KB $\tilde{K}$, retriever Retrieve, verifier Verify, weight map W, subclaim c

function VerifySubclaim($c, \tilde{K}$, Retrieve, Verify, W)

$(E, \rho) \leftarrow \text{Retrieve}(c, \tilde{K}, W)$

$(S, F, U) \leftarrow \text{Verify}(E, c)$

$s^+ \leftarrow \sum_{e \in S} \rho(e) \cdot \text{GetWeight}(W, e)$

$s^- \leftarrow \sum_{e \in F} \rho(e) \cdot \text{GetWeight}(W, e)$

$d \leftarrow \begin{cases} +1, & \text{if } s^+ \geq s^- \\ -1, & \text{otherwise} \end{cases}$

$m \leftarrow |s^+ - s^-|$

for all $e \in S \cup F$

$\sigma(e) \leftarrow \begin{cases} +1, & \text{if } e \in S \text{ (supporting)} \\ -1, & \text{if } e \in F \text{ (refuting)} \end{cases}$

UpdateWeight$(e, d, \sigma(e), m)$

return d

We instead use structured dataset annotations, including gold evidence, entity identifiers, and truth labels, to generate topically aligned misinformation. For **true** claims, we prompt an LLM with the claim, gold evidence, and relevant entities and ask it to reuse those entities in a coherent but misleading context that contradicts the gold evidence. For **false** claims, we prompt the LLM to fabricate plausible evidence that supports the false claim using the same relevant entities. This procedure produces fluent, claim-relevant distractors that are more likely to be retrieved and more difficult for standard verification pipelines to dismiss.

This construction serves as a controlled adversarial stress test for retrieval-based fact checking rather than a complete simulation of real-world misinformation. To complement it, we also evaluate on RAGuard (Zeng et al., 2025), which contains naturally occurring political and social-media misinformation. Unlike our synthetic misinformation, which is explicitly constructed against claim-specific gold evidence, RAGuard misinformation is not targeted toward a particular gold passage. In our controlled setting, synthetic misinformation is therefore frequently retrieved together with its corresponding gold evidence: in more than 90% of cases where the gold evidence is retrieved, its misinformation counterpart is retrieved as well. Across datasets, misinformation can account for a substantial portion of retrieved evidence and often exceeds the amount of retrieved gold evidence, creating a deliberately challenging verification setting.

3.3 Claim Decomposition

Prior work such as GraphCheck (Jeon & Lee, 2025) has shown that decomposing complex claims into simpler subcomponents can improve multi-hop fact checking. Following this approach, DEDUCE uses an LLM to decompose each original claim into a set of simpler subclaims. Claim decomposition itself is not introduced as a new component of DEDUCE; instead, we use it to expose the verifier

to multiple evidence neighborhoods under contaminated retrieval. The decomposition reduces the reasoning burden of each verification step and allows the retriever to search for evidence directly related to individual factual components.

Algorithm 2 Full Claim Verification

Input: contaminated KB $\tilde{K}$,
retriever Retrieve, verifier Verify, weight map W, claim c

```
subclaims ← DECOMPOSE(c)
for c_s in subclaims
    if not VERIFYSUBCLAIM(c_s, K̃, Retrieve, Verify, W)
        return False
return True
```

Claim decomposition is especially useful under contaminated retrieval. Different subclaims may retrieve different evidence neighborhoods, making verification less dependent on any single noisy retrieval result. In dense corpora, true claims are often supported by multiple related passages, and this redundancy can provide a useful signal even when some retrieved documents are misleading. Claim decomposition therefore provides diverse evidence observations on which the subsequent stance aggregation and reliability estimation operate. We provide the prompt for claim decomposition in Appendix G.4.

3.4 Fact Checking With Document Weight Updating

We maintain a reliability state for each document in a weight map W. Each document entry contains two statistics, pos and neg, both initialized to 0. The pos statistic accumulates evidence that the document's stance agrees with aggregate verification decisions, whereas neg accumulates evidence that it conflicts with them. When accessed, these statistics are converted into a reliability value through the sigmoid transformation in Algorithm 3, yielding a value in $[0, 1]$. Consequently, an unseen document initially has reliability 0.5.

In our default setting, W is initialized once for a KB and maintained across claims processed against that KB, allowing reliability evidence to accumulate online as verification proceeds. Thus, the default protocol uses a persistent passage-level reliability state rather than resetting document weights independently for every claim. We additionally evaluate within-claim updates, per-claim resets, and a held-out calibration setting with frozen weights in Section 4.

For a given big and complicated claim, we first decompose it into simpler subclaims. For each subclaim, we use it as query to the retriever, and then the retriever returns an evidence set E with relevance scores $\rho(e)$ for each document e that are related to the subclaim. The LLM verifier partitions E into 3 different stance groups: supporting, refuting, and unrelated, denoted by S, F, and U, separately. Next, we compute the aggregated weighted support and refutation scores, s^+ and s^-, by combining retrieval relevance with the current document reliability:

Algorithm 3 Access and Update Weight

Input: weight map W, evidence e, prediction d, stance σ, update magnitude m

```
function UPDATEWEIGHT(e, d, σ, m)
    if d = σ
        W[e].pos ← W[e].pos + m
    else
        W[e].neg ← W[e].neg + m

function GETWEIGHT(W, e)
    p, n ← W[e].pos, W[e].neg
    return 1 / (1 + exp(n − p))
```

$$s^+ = \sum_{e \in S} \rho(e)\, \textsc{GetWeight}(W, e), \quad s^- = \sum_{e \in F} \rho(e)\, \textsc{GetWeight}(W, e).$$

The subclaim prediction is $d \in \{+1, -1\}$, where $+1$ denotes true and -1 denotes false. We set $d = +1$ if $s^+ \geq s^-$ and $d = -1$ otherwise. For each document, $\sigma(e)$ denotes its stance sign, with $+1$ for support and -1 for refutation.

After the subclaim decision is made, documents in the supporting and refuting groups are updated according to whether their stance agrees with the aggregate decision. If $\sigma(e) = d$, the document's pos statistic is increased; otherwise, its neg statistic is increased. We use the decision margin $m = |s^+ - s^-|$ as the update magnitude, so higher-margin subclaim decisions induce stronger reliability updates. Unrelated documents are not updated.

The resulting reliability estimates affect not only stance aggregation but also subsequent evidence retrieval. When candidate documents are retrieved for a later subclaim, the retriever incorporates their current reliability values into candidate reranking, promoting documents that repeatedly agree

with aggregate verification decisions and demoting those that repeatedly conflict with them. This retrieval-feedback step closes the loop between verification and evidence selection: previous stance observations affect document reliability, which in turn affects which evidence is prioritized in future retrievals. This distinguishes DEDUCE's use of reliability estimation from a static post-retrieval aggregation procedure.

Finally, the original claim is classified as TRUE if and only if all of its subclaims are identified as true, as shown in Algorithm 2. This strict conjunction provides a conservative verification rule under contaminated evidence. We analyze its false-negative behavior in Section 4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Table 1: Macro-F1 under clean and contaminated retrieval. Misinfo Baseline denotes the decomposition-enabled baseline under contaminated retrieval, while Misinfo Weight Update denotes DEDUCE.

Datasets	Mode	Qwen3 14B	Gemma-3 12B	Llama-3.1 8B	Mistral-7B v0.3
HoVer	Clean	66.15	63.58	66.44	61.03
	Misinfo Baseline	50.15	48.86	42.34	49.86
	Misinfo Weight Update	58.33 (+8.18)	57.10 (+8.24)	55.43 (+13.09)	56.05 (+6.19)
EX-FEVER	Clean	68.43	62.20	63.06	64.29
	Misinfo Baseline	46.28	47.39	44.74	49.69
	Misinfo Weight Update	58.71 (+12.43)	60.48 (+13.09)	58.63 (+13.89)	58.26 (+8.57)
HotpotQA	Clean	72.18	71.07	67.05	67.69
	Misinfo Baseline	48.71	48.49	49.03	43.66
	Misinfo Weight Update	59.93 (+11.22)	56.43 (+7.94)	60.12 (+11.09)	54.64 (+10.98)
PUBHEALTH	Clean	64.23	54.86	60.91	56.99
	Misinfo Baseline	52.76	48.66	52.68	51.39
	Misinfo Weight Update	60.06 (+7.30)	52.63 (+3.97)	58.52 (+5.84)	55.06 (+3.67)
SciFact	Clean	63.89	62.38	69.48	73.41
	Misinfo Baseline	49.11	44.13	43.37	49.32
	Misinfo Weight Update	61.78 (+12.67)	60.80 (+16.67)	63.76 (+20.39)	64.36 (+15.04)
Climate-FEVER	Clean	66.26	62.54	64.75	65.93
	Misinfo Baseline	34.82	28.98	31.29	37.93
	Misinfo Weight Update	49.28 (+14.46)	46.01 (+17.03)	56.78 (+25.49)	51.27 (+13.34)
RAGuard	Clean	62.59	63.24	57.73	60.05
	Misinfo Baseline	48.71	41.45	48.99	48.23
	Misinfo Weight Update	56.86 (+8.15)	54.79 (+13.34)	53.55 (+4.56)	55.74 (+7.51)

Datasets and Knowledge Sources. We evaluate DEDUCE on seven benchmarks: HoVer (Jiang et al., 2020), EX-FEVER (Ma et al., 2023), HotpotQA (Yang et al., 2018), PubHealth (Kotonya & Toni, 2020), SciFact (Wadden et al., 2020), Climate-FEVER (Diggelmann et al., 2020), and RAGuard (Zeng et al., 2025). HotpotQA is converted into declarative fact-checking claims. RAGuard contains naturally occurring supporting, misleading, and unrelated evidence; unlike our targeted synthetic misinformation, its misleading evidence is untargeted, which helps explain why gold retrieval exceeds misinformation retrieval in Table 6. Additional details are provided in Appendix A.

Baselines and Implementation. Our primary comparison includes a clean/reference condition, a decomposition-enabled contaminated baseline, and full DEDUCE. We additionally compare with GraphCheck (Jeon & Lee, 2025), ProgramFC (Pan et al., 2023), PACAR (Zhao et al., 2024), LoCal (Ma et al., 2025), RobustRAG (Xiang et al., 2024), TrustRAG (Zhou et al., 2025), Astute RAG (Wang et al., 2025), TruthFinder (Yin et al., 2007), Investment (Pasternack & Roth, 2010), and AccuSim (Dong et al., 2009). We use Qwen3-14B, Gemma-3-12B-it, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 as verifier backbones. Retrieval combines BM25 (Robertson et al., 2009) and FAISS (Douze et al., 2024) with *e5-base-v2*. All decomposition-enabled methods use the same precomputed subclaims generated by Qwen2.5-72B-Instruct.

4.2 MAIN RESULTS

Fact-Checking under Contamination. Table 1 shows substantial degradation under contamination across all datasets and backbones, while DEDUCE consistently recovers part of the loss. For example, Qwen3-14B on Climate-FEVER drops from 66.26 to 34.82 and recovers to 49.28 with

DEDUCE. GraphCheck and ProgramFC exhibit similar degradation (Table 7), confirming that FCC is challenging beyond our baseline.

Table 2: Stance-classification Macro-F1 on misinformation.

Dataset	Qwen3	Gemma-3	Llama-3.1	Mistral-7B
HoVer	57.86	59.30	49.59	58.53
EX-FEVER	44.99	55.25	47.24	54.73
HotpotQA	54.97	59.21	51.36	54.94
PUBHEALTH	61.60	34.10	50.45	67.29
SciFact	51.27	46.64	48.79	49.31
Climate-FEVER	66.88	68.39	63.33	65.99
RAGuard	79.00	80.90	83.05	87.24

Table 3: Recent fact-checking methods on HoVer with Qwen3-14B.

Method	Clean	Contam.	Drop
PACAR (Zhao et al., 2024)	67.87	47.92	19.95
LoCal (Ma et al., 2025)	66.95	45.26	21.69
DEDUCE	66.15	**58.33**	**7.82**

Table 4: Comparison with robust-RAG defenses using Qwen3-14B.

Method	HoVer	EX-FEVER
RobustRAG (Xiang et al., 2024)	55.42	56.14
TrustRAG (Zhou et al., 2025)	52.96	57.58
Astute RAG (Wang et al., 2025)	57.03	54.25
DEDUCE	58.33	58.71
DEDUCE + RobustRAG	**59.67**	**60.19**

Table 5: Effectiveness of misinformation generation.

Method	Verification F1		Retrieval Count	
	HoVer	EX-FEVER	HoVer	EX-FEVER
Clean	66.15	68.43	338770	368010
Keyword Replacement	57.07	55.64	46744	30133
Direct Negation	59.22	58.93	42113	28965
F3 (Lucas et al., 2023)	52.76	51.91	69338	47023
Our Generation	**50.15**	**46.28**	**89012**	**53671**

Recent Fact-Checking Methods. PACAR and LoCal slightly outperform DEDUCE on clean evidence but drop by 19.95 and 21.69 points under contamination, compared with 7.82 for DEDUCE (Table 3), indicating that strong reasoning alone does not ensure robustness to corrupted retrieval.

Robust-RAG Defenses. RobustRAG uses isolate-then-aggregate, TrustRAG applies cluster filtering and consistency resolution, while Astute RAG additionally uses parametric model knowledge. DEDUCE outperforms all three on both datasets. Combining RobustRAG with DEDUCE further reaches 59.67/60.19, showing that filtering and reliability-aware reranking are complementary. Because Astute RAG uses an additional parametric-knowledge channel, its comparison is not fully information-matched.

Misinformation Generation. Our generation procedure causes the largest verification degradation while achieving the highest retrieval frequency (Table 5). An independent model further evaluates a stratified random 30% sample of synthesized evidence, and every sampled passage contains at least one false statement.

4.3 Ablation and Robustness Analysis

Table 6: Retrieval statistics across datasets and verifier backbones.

Dataset	Type	Qwen3	Gemma-3	Llama-3.1	Mistral-7B
HoVer	Total	338770	348750	346790	327560
	Misinfo	89012 (26.28%)	90104 (25.84%)	88525 (25.53%)	100439 (30.66%)
	Golden	33795 (9.98%)	31649 (9.07%)	30841 (8.89%)	32289 (9.86%)
EX-FEVER	Total	368010	376820	375350	365810
	Misinfo	53671 (14.58%)	52053 (13.81%)	51253 (13.65%)	57895 (15.83%)
	Golden	27843 (7.57%)	24986 (6.63%)	25100 (6.69%)	26999 (7.38%)
HotpotQA	Total	442380	450150	423030	416700
	Misinfo	67723 (15.31%)	69144 (15.36%)	68766 (16.26%)	76019 (18.24%)
	Golden	37967 (8.58%)	36096 (8.02%)	36124 (8.54%)	37493 (9.00%)
PUBHEALTH	Total	24170	25110	26760	21540
	Misinfo	6618 (27.38%)	7251 (28.88%)	7303 (27.29%)	6754 (31.36%)
	Golden	2107 (8.72%)	2197 (8.75%)	2301 (8.60%)	1998 (9.28%)
SciFact	Total	31530	31040	28290	29020
	Misinfo	6151 (19.51%)	6135 (19.76%)	6099 (21.56%)	5875 (20.24%)
	Golden	2656 (8.42%)	2610 (8.41%)	2265 (8.01%)	2525 (11.72%)
Climate-FEVER	Total	84200	82680	76470	79590
	Misinfo	30214 (35.88%)	28765 (34.79%)	26328 (34.43%)	30250 (38.01%)
	Golden	3800 (4.51%)	3540 (4.28%)	3124 (4.09%)	3637 (4.57%)
RAGuard	Total	5483	7591	7283	6011
	Misinfo	1172 (21.38%)	1228 (16.18%)	1151 (15.80%)	910 (15.14%)
	Golden	1789 (32.63%)	1900 (25.03%)	1739 (23.88%)	1369 (22.77%)

Table 7: Macro-F1 of existing fact-checking frameworks under clean and contaminated retrieval.

Dataset	GraphCheck Clean	GraphCheck Misinfo	ProgramFC Clean	ProgramFC Misinfo
HoVer	65.15	51.21	61.60	46.63
EX-FEVER	64.92	45.78	66.40	42.82
HotpotQA	70.14	49.71	72.48	27.28
PUBHEALTH	63.77	50.73	63.48	50.87
SciFact	65.49	49.86	70.69	42.97
Climate-FEVER	61.06	37.74	63.06	32.94
RAGuard	62.13	43.09	61.77	40.58

Table 8: Sequential component ablation with Qwen3-14B.

Configuration	Decomp.	Stance	Update	HoVer	EX-FEVER
Vanilla contaminated RAG	No	No	No	42.36	39.87
RAG + decomposition	Yes	No	No	50.15	46.28
DEDUCE, stance only	Yes	Yes	No	55.02	51.24
DEDUCE (full)	Yes	Yes	Yes	**58.33**	**58.71**

Table 9: Comparison with classical truth-discovery methods.

Method	Granularity	Retrieval	HoVer	EX-FEVER
TruthFinder (Yin et al., 2007)	Source	Static	56.92	57.62
Investment (Pasternack & Roth, 2010)	Source	Static	55.45	56.16
AccuSim (Dong et al., 2009)	Source	Static	52.13	53.43
DEDUCE (no feedback)	Passage	Query-dep.	53.29	54.87
DEDUCE (full)	Passage	Query-dep.+feedback	**58.33**	**58.71**

Component Ablation. Because the contaminated baseline in Table 1 already uses decomposition, we start from vanilla RAG and add components sequentially. Decomposition contributes +7.79/+6.41 points on HoVer/EX-FEVER, stance grouping adds +4.87/+4.96, and reliability updating adds +3.31/+7.47. Thus, all three components contribute substantially.

Classical Truth Discovery. Full DEDUCE exceeds TruthFinder by 1.41/1.09 points. More notably, removing retrieval feedback reduces performance from 58.33/58.71 to 53.29/54.87, showing that the retrieval-feedback loop is more important than the reliability update rule in isolation.

Table 10: Ablation on document reliability updating.

Qwen3-14B			
Dataset	**No Update**	**Linear**	**Weighted**
HoVer	55.02	56.09 (+1.07)	58.33 (+3.31)
EX-FEVER	51.24	53.77 (+2.53)	58.71 (+7.47)
Llama-3.1-8B			
Dataset	**No Update**	**Linear**	**Weighted**
HoVer	50.83	52.96 (+2.13)	55.43 (+4.60)
EX-FEVER	52.29	56.07 (+3.78)	58.63 (+6.34)

Table 11: Higher misinformation ratios on HoVer with Qwen3-14B.

Ratio	**Misinfo**	**Gold**	**Base**	**Ours**
1/4	89012	33795	50.15	58.33
1/3	122136	32684	42.16	49.68
1/2	178491	28706	23.07	27.44

Table 12: Performance under reduced misinformation contamination.

Dataset	**Drop**	**Total**	**Misinfo**	**Gold**	**Base**	**Ours**
Qwen3-14B						
HoVer	0%	338770	89012 (26.28%)	33795 (9.98%)	50.15	58.33 (+8.18)
	30%	330610	60573 (18.32%)	31508 (9.53%)	50.48	59.72 (+9.24)
	50%	331570	45023 (13.58%)	31205 (9.41%)	53.15	60.88 (+7.73)
	70%	329040	35274 (10.72%)	31922 (9.70%)	57.29	62.78 (+5.49)
EX-FEVER	0%	368010	53671 (14.58%)	27843 (7.57%)	46.28	58.71 (+12.43)
	30%	363860	33386 (9.18%)	25652 (7.05%)	52.32	61.78 (+9.46)
	50%	352470	15935 (4.90%)	24860 (7.05%)	54.86	62.76 (+7.90)
	70%	359440	8904 (2.47%)	25134 (6.99%)	59.16	64.91 (+5.75)
Llama-3.1-8B-Instruct						
HoVer	0%	346790	88525 (25.53%)	30841 (8.89%)	42.34	55.43 (+13.09)
	30%	324930	56309 (17.33%)	30398 (9.36%)	45.47	58.27 (+12.80)
	50%	339630	38391 (11.30%)	30219 (8.90%)	48.89	61.65 (+12.76)
	70%	334970	26104 (7.79%)	30263 (9.03%)	52.04	63.53 (+11.49)
EX-FEVER	0%	375350	51253 (13.65%)	25100 (6.69%)	44.74	58.63 (+13.89)
	30%	368320	31739 (8.62%)	24840 (6.74%)	51.42	59.74 (+8.32)
	50%	367980	18945 (5.15%)	24476 (6.65%)	52.45	61.06 (+8.61)
	70%	365730	10231 (2.80%)	24965 (6.83%)	56.74	62.13 (+5.39)

Table 13: Reliability persistence on HoVer with Qwen3-14B.

Protocol	**Macro-F1**
Per-claim reset	55.40
Within-claim only	56.60
Cross-claim persistent	**58.33**
Held-out calibration	57.50

Table 14: Effect of additional reweighting iterations on HoVer.

Iter.	**Qwen3**	**Gemma-3**	**Llama-3.1**	**Mistral-7B**
1	58.33	57.10	55.43	56.05
2	58.56	57.59	55.82	56.33
3	59.04	58.31	56.14	57.07

Table 15: Error analysis of strict subclaim conjunction with Qwen3-14B.

Dataset	**FN**	**Single FN**	**TP**	**Recall**
HoVer	765	191	1235	61.75
EX-FEVER	448	118	1557	77.66
HotpotQA	1134	249	2549	69.21

Weight Updating Strategies. The full weighted strategy performs best across both datasets and backbones.

Contamination Level. DEDUCE consistently outperforms the baseline when misinformation is progressively removed (Table 12).

Higher Misinformation Ratios. Both methods deteriorate as misinformation becomes dominant, while DEDUCE remains stronger. The shrinking improvement at the highest ratio exposes the limitation of consistency-based reliability when truthful evidence becomes sparse.

Online Reliability and Evidence Recurrence. The default setting accumulates reliability across claims. We compare alternative protocols in Table 13.

Passage reuse is common: 34% recur across subclaims of the same claim and 61% across different claims, with mean reuse 3.4. Persistent updating performs best, while held-out frozen calibration still reaches 57.50, showing that most of the gain does not require test-stream adaptation.

Reweighting Iterations and Claim Order. Additional iterations provide only modest gains (Table 14). Across 10 shuffled claim orders, the largest standard deviation is 0.007 Macro-F1 and the largest range is 0.023, indicating limited order sensitivity.

Coordinated Misinformation. We add two lexically distinct paraphrases of each misinformation passage, creating three mutually consistent false passages. Performance drops from 58.33 to 39.45 on HoVer and from 58.71 to 42.66 on EX-FEVER, showing that stance consistency alone is insufficient against coordinated false narratives.

Adaptive Reputation Poisoning. When previously reliable documents are modified to support false targets while retaining their weights, attack success rises to 65% and Macro-F1 falls to 43.07. Resetting their weights improves F1 to 48.33 and reduces attack success to 57%, suggesting content-version tracking as a natural mitigation.

Weight Distribution and Mixed-Truth Passages. Figure 2 shows that gold evidence generally receives higher weights than misinformation. However, approximately 10% of misinformation passages are incorrectly upweighted when passages contain mixed truthful and misleading content (Appendix E).

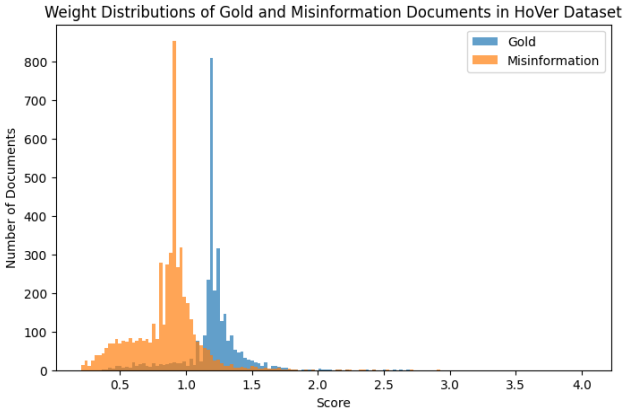


Figure 2: Final weight distributions for truth and misinformation on HoVer.

Strict Subclaim Conjunction. Single-reject cases account for only a subset of false negatives (Table 15), indicating that strict conjunction contributes to recall loss but is not its dominant cause.

Efficiency. The main cost scales approximately linearly with the number of subclaims. Across all datasets, claims contain 2.6 subclaims on average (median 3; 95th percentile 5), while stance aggregation and reliability updates are lightweight. Detailed runtime and memory statistics are provided in Appendix F.2.

5 Conclusion

We study Fact-Checking under Contamination (FCC), a setting in which truthful and misleading documents coexist in retrieved evidence and may jointly influence a verifier's decision. To address this challenge, DEDUCE combines claim decomposition, LLM-based stance aggregation, passage-level reliability estimation, and retrieval-feedback reranking into a unified framework. We further introduce a controlled contamination protocol that generates topically aligned adversarial evidence, and complement this setting with naturally occurring misinformation to evaluate robustness under more realistic conditions. Across seven benchmarks and four open-source LLM backbones, we find that contamination substantially degrades standard verification pipelines, while DEDUCE consistently recovers a substantial portion of the lost performance. Our results also show that DEDUCE outperforms recent fact-checking methods under contaminated retrieval, as well as adaptations of classical truth-discovery methods and purpose-built robust-RAG defenses on HoVer and EX-FEVER. Ablation studies confirm that claim decomposition, stance grouping, and reliability updating each contribute to the overall gains. Moreover, feeding learned reliability back into retrieval provides an additional benefit beyond static reliability aggregation, suggesting that evidence assessment and evidence selection can reinforce one another. Despite these improvements, our results reveal important limitations. Coordinated misinformation and adaptive reputation poisoning remain challenging, particularly when misleading evidence is highly redundant, mutually consistent, or inherits reliability accumulated from earlier interactions. In such cases, text-based stance consistency alone may not provide enough information to distinguish trustworthy evidence from coordinated or strategically amplified claims. These findings highlight both the value and the limits of explicitly modeling evidence reliability in retrieval-augmented fact checking. Future work could incorporate complementary signalsâĂŤincluding source provenance and independence, temporal information, and content-version trackingâĂŤto better detect coordinated misinformation and maintain reliable evidence assessments as the knowledge environment evolves.

AI USE STATEMENT

We used generative AI tools to check grammar, refine sentence structure, and improve the clarity and readability of the writing. Their use was limited to language editing, with the aim of communicating our ideas more clearly while preserving the intended meaning and technical content. The authors retain full responsibility for the content and accuracy of the manuscript.

ETHICS STATEMENT

This work aims to improve the robustness of automated fact-checking when retrieved evidence contains misinformation. Our experiments use existing benchmarks and synthetically generated misleading passages for controlled robustness evaluation. We acknowledge the dual-use risks of the misinformation construction procedure, which could be repurposed to produce plausible false evidence or manipulate retrieval-based systems. Such synthetic passages should be clearly identified as research artifacts and should not be presented as factual sources. DEDUCE may also inherit biases from its underlying language models and evidence corpora, and its reliability estimates reflect agreement among retrieved passages rather than independently established truth. As our stress tests demonstrate, coordinated misinformation and reputation poisoning can reinforce incorrect predictions. Accordingly, the framework should support human judgment, with independent evidence validation and human oversight in consequential applications.

REPRODUCIBILITY STATEMENT

We document the DEDUCE framework in Section 3, with pseudocode for subclaim verification, claim-level prediction, and document reliability updating in Algorithms 1–3. Section 4.1 and Appendix A describe the datasets, preprocessing procedures, verifier backbones, and retrieval configuration. Section 3.2 describes the synthetic misinformation construction procedure, while Appendix G provides the prompts for evidence generation, stance classification, and claim decomposition. Appendix B further explains the decomposition procedure, and Appendix D reports the hardware configuration, memory usage, and inference efficiency. We explicitly describe the persistent cross-claim reliability state in Section 3.4 and examine alternative update protocols, repeated iterations, and sensitivity to claim order in Section 4.3. These details support reproduction of the experimental pipeline and assessment of the factors affecting the reported results.

REFERENCES

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009. doi: 10.14778/1687627.1687690.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K. Bressem, Jakob Nikolas Kather, and Daniel Truhn. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7:288, 2024. doi: 10.1038/s41746-024-01282-7.

Hyewon Jeon and Jay-Yoon Lee. Graphcheck: Multi-path fact-checking with entity-relationship graphs. *arXiv preprint arXiv:2502.20785*, 2025.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*, 2020.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Jason S Lucas, Adaku Uchendu, Michiharu Yamashita, Joo-Young Lee, Shaurya Rohatgi, and Dongwon Lee. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14279–14305, 2023.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*, 2023.

Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. LoCal: Logical and causal fact-checking with llm-based multi-agents. In *Proceedings of the ACM on Web Conference 2025*, pp. 1614–1625. Association for Computing Machinery, 2025. doi: 10.1145/3696410.3714748.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*, 2023.

Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 877–885, Beijing, China, 2010. Coling 2010 Organizing Committee.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1476.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. Defending against disinformation attacks in open-domain question answering. *arXiv preprint arXiv:2212.10002*, 2022.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1048–1052, 2007. doi: 10.1145/1281192.1281309.

Linda Zeng, Rithwik Gupta, Divij Motwani, Yi Zhang, and Diji Yang. Worse than zero-shot? a fact-checking dataset for evaluating the robustness of rag against misleading retrievals. *arXiv preprint arXiv:2502.16101*, 2025.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12564–12573, Torino, Italia, 2024. ELRA and ICCL.

Hongjian Zhou, Xinyu Zou, Jinge Wu, Sean Wu, Junchi Yu, Bradley Max Segal, Tobias Erich Niebuhr, Sara Amro, Michael Petrus, Sheikh Momin, Alexandra M. Cardoso Pinto, Rachel Niesen, Laura Sophie Wegner, Dhruv Darji, Jung Moses Koo, Joshua Fieggen, Kapil Narain, Mingde Zeng, Lei Clifton, Linda Shapiro, Fenglin Liu, and David A. Clifton. Measuring epistemic resilience of llms under misleading medical context. *arXiv preprint arXiv:2606.12291*, 2026.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. Trustrag: Enhancing robustness and trustworthiness in retrieval-augmented generation. *arXiv preprint arXiv:2501.00879*, 2025.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, Seattle, WA, 2025. USENIX Association.

A Experimental Details

Dataset Preprocessing and Statistics The datasets used in our evaluation required specific preprocessing steps to align with the fact-checking task. For **HotpotQA**, we transformed the original multi-hop questions into declarative statements by combining the question and its answer. We then prompted a large language model to generate veracity-conditioned claims based on the provided gold evidence to create a balanced set of true and false statements. For **RAGuard**, we utilized a corpus of 2,648 political claims from the 2000–2024 U.S. presidential cycles, grounded in a knowledge base of 16,331 documents scraped from Reddit. This dataset provides a critical testbed for naturally occurring, noisy evidence that differs significantly from curated sources like Wikipedia.

Contaminated Corpora Construction. To simulate realistic adversarial environments, we construct contaminated versions of each KB. For every claim, we inject two types of distractors: (1) *Refutations* that contradict gold evidence, and (2) *Fake Evidence* that falsely supports incorrect claims. This setup ensures that misinformation is topically aligned with the query, challenging standard retrieval mechanisms.

Retrieval Configuration Our retrieval pipeline is designed to maximize recall by leveraging both lexical and semantic search indices for every corpus. We utilize a **BM25** lexical index tuned with hyperparameters $k_1 = 0.9$ and $b = 0.4$ to ensure strong performance on exact keyword matches. This is complemented by a dense **FAISS** index using the `intfloat/e5-base-v2` embedding model to capture semantic similarities that lexical search might miss. To combine these approaches, we employ a hybrid search strategy using a reciprocal rank fusion method with a balancing parameter of $\alpha = 0.5$. This configuration retrieves the top $K = 10$ documents for each subclaim, allowing the framework to anchor on exact matches while remaining sensitive to paraphrased or semantically related misinformation.

Prompting and Backbone Settings The technical implementation of DEDUCE relies on high-quality initial processing and robust inference. For the primary task of claim-to-subclaim decomposition, we follow established best practices by employing **Qwen2.5-72B-Instruct**, which ensures high logical granularity and consistency in the generated sub-components. All subsequent verifier backbone models, including Qwen3, Llama, Mistral, and Gemma variants, were executed using `bfloat16` precision on high-performance computing clusters. We provide the full text of the

Table 16: Computation efficiency per example for each dataset and each LLM we used. All experiments are conducted on a single Nvidia H200 GPU. The memory usages for Qwen3-14B, Gemma-3-12b-it, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 are 128 GB, 120 GB, 127 GB, 127 GB, respectively. We use huggingface models and the vLLM Kwon et al. (2023) acceleration.

Misinfo Datasets	**Qwen3 14B**	**Gemma-3 12b-it**	**Llama-3.1 8B-Instruct**	**Mistral-7B Instruct-v0.3**
HoVer	1.47s	1.62s	1.09s	1.18s
EX-FEVER	3.05s	3.73s	2.05s	2.43s
HotpotQA	2.36s	2.16s	1.34s	1.55s
PUBHEALTH	1.44s	1.21s	0.83s	0.97s
SciFact	2.09s	2.01s	1.27s	1.34s
Climate-FEVER	2.42s	2.36s	1.34s	1.58s
RAGuard	1.86s	1.75s	1.24s	1.47s

prompts used for stance detection, evidence synthesis, and the iterative weight feedback loops in Section G to ensure complete reproducibility of our verification results.

B GraphCheck Claim Decomposition

As suggested in Section 3, we believe that the redundancy of truthful information and the density of subclaims contribute to the robustness of DEDUCE. Accordingly, instead of prompting directly on the original claim, DEDUCE first leverages an LLM to decompose a complicated claim into multiple simple subclaims. The original claim is predicted as true if and only if all subclaims are predicted as true.

Sometimes, the original claim contains implicit references, and graph decomposition also helps fill in those implicit entities. For more technical details, we refer the reader to the original GraphCheck paper (Jeon & Lee, 2025). Within DEDUCE, the additional subclaims generated by graph decomposition produce denser queries, bringing in more diverse support from non-gold but informative documents in the contaminated knowledge base.

We use Qwen2.5-72B-Instruct as base backbone to perform graph decomposition as suggested in the GraphCheck paper. We found that with graph decomposition, more documents were involved in the fact checking process, and thus provide more diverse and redundant support towards the truth. The decomposition can be performed in advance, and no need to perform every time for fact checking.

C Error Analysis for Document Classification

Because the clean knowledge base is indexed at the passage level, we treat misinformation at the same granularity; sentence-level segmentation would sacrifice the local context needed for reliable stance assessment. Under this convention, a passage is labeled as misinformation if it contains at least one materially false statement. However, mixed-truth passages typically also contain numerous true sentences. During iterative inference, these truths can spuriously bolster support for some subclaims, causing the passage to be upweighted, and thereby elevating the false-negative rate.

D Computation Efficiency Report

We report the computation efficiency per example for each dataset and each LLM we used (See Table 16).

E Additional Discussion on Mixed-Truth Passages

DEDUCE performs document weighting at the passage level. This choice preserves semantic coherence and provides sufficient local context for stance classification. However, passage-level modeling can introduce a mixed-truth risk: a passage may contain both correct and incorrect statements. Since

we label a passage as misinformation if it contains at least one materially false statement, some misinformation passages may still include true statements that support certain subclaims.

This creates a possible failure mode for dynamic reweighting. If a mixed-truth passage supports a true subclaim through its correct content, it may receive a higher reliability weight even though other parts of the passage are false. This is a limitation of passage-level evidence modeling. Sentence-level evidence modeling could reduce this risk, but it may also remove useful context needed for reliable stance assessment.

In practice, we observe that this risk is limited. Misinformation documents that spuriously support subclaims are often classified as unrelated rather than supporting, which prevents systematic reinforcement. We also find that the false-positive rate of misinformation being incorrectly upweighted is low, around 10%. Therefore, while mixed-truth passages remain a potential source of error, they do not dominate the behavior of the reweighting mechanism in our experiments.

F Additional Discussions

F.1 Additional Discussion on Coordinated Misinformation and Low-Redundancy Settings

DEDUCE relies on aggregate stance consistency across retrieved evidence. This design is motivated by the observation that, in large knowledge bases, truthful information is often repeated across multiple documents, while misleading evidence is less likely to be both dominant and mutually consistent. However, this assumption can fail under coordinated misinformation. If many false documents are self-consistent and numerically dominate the retrieved evidence pool, a majority-based reweighting mechanism within DEDUCE may incorrectly reinforce the false narrative.

This limitation is closely related to echo-chamber settings, where many sources repeat the same misleading claim. In such cases, the available evidence may no longer provide a reliable signal for distinguishing truth from misinformation. DEDUCE higher-misinformation-ratio analysis in Section 4 partially reflects this issue: as misinformation becomes increasingly dominant, both the baseline and DEDUCE degrade substantially. Nevertheless, DEDUCE remains consistently stronger than the baseline, indicating that reweighting still provides some robustness before misinformation fully overwhelms truthful evidence.

Low-redundancy domains present another challenge. In professional or scientific domains such as SciFact, evidence for a claim may be sparse, and different claims may rely on largely independent documents. In this case, an early verifier error has less opportunity to be corrected by repeated agreement across many related claims. Since our weight updates are aggregated across many examples, a single mistaken update is unlikely to dominate the whole corpus, but low evidence redundancy still limits the strength of the reliability signal. This suggests that future work could combine dynamic reweighting with uncertainty estimation or source-level credibility modeling for sparse professional domains.

F.2 Additional Discussion on Efficiency and Computational Overhead

The main computational overhead of DEDUCE comes from claim decomposition and subclaim-level retrieval and verification. If an original claim is decomposed into n subclaims, DEDUCE performs retrieval and verification for each subclaim. Therefore, the cost scales approximately linearly with the number of generated subclaims. In practice, the number of reasoning hops is usually small, so the additional cost remains manageable.

The document reweighting operation itself is lightweight. It only requires aggregating support and refutation scores and updating scalar reliability statistics for retrieved documents. Thus, the cost of weight updating is small compared with LLM-based verification. Our iteration analysis further shows that most gains are achieved after one reweighting iteration, while additional iterations provide only marginal improvements. Therefore, we use one iteration by default to balance robustness and efficiency.

The claim decomposition step can also be performed in advance for static benchmark evaluation. This further reduces repeated inference overhead when the same claims are evaluated under different retrieval or weighting settings.

G PROMPTS

G.1 VERIFIER: SUBCLAIM STANCE GROUPING PROMPT

```
You are an expert at distinguish statements's agreements or attitude
towards a claim.
You will be given a list of statements and a claim. Then you need you
identify the attitude for each statement towards the given claim.
You are going to group these statements into exactly three groups:
SUPPORT, REFUTE, NOT_RELATED.
SUPPORT means: the statement agrees with the claim.
REFUTE means: the statement does not agree with the claim.
NOT_RELATED means: the statement has nothing to do with the claim.

Here are some examples:

Example 1:
# Statements:
    0. Charles James Eastwood, better known as Jim Eastwood, is a
    Northern Irish Businessman and formerly one of the final four
    contestants in the seventh UK series of "The Apprentice". He was
    born in Cookstown, County Tyrone, Northern Ireland and is a
    graduate of the University of Ulster having also attended Harvard
    for a two-week course and the University of North Carolina.
    During his time on The Apprentice, he gained the nickname "Jedi
    Jim" due to his persuasive abilities and use of mind
    games.,Eastwood is also a former all-Ireland cycling champion.

    1. Cookstown is a town and townland in County Tyrone, Northern
    Ireland. It is the fourth largest town in the county and had a
    population of nearly 11,000 people in the 2001 Census. It is one
    of the main towns in the area of Mid-Ulster. It was founded
    around 1620 when the townlands in the area were leased by an
    English ecclesiastical lawyer, Dr. Alan Cooke, from the
    Archbishop of Armagh, who had been granted the lands after the
    Flight of the Earls during the Plantation of Ulster. It was one
    of the main centres of the linen industry West of the River Bann,
    and until 1956, the processes of flax spinning, weaving,
    bleaching and beetling were carried out in the town.

    2. Robert Ross (Roy) Knight (12 December 1891 - 11 September
    1971) was a Co-operative Commonwealth Federation member of the
    Canadian House of Commons. He was born in Cookstown, County
    Tyrone, Northern Ireland and became a farmer and teacher by
    career.

    3. Cookstown is a coastal fishing village in southern Ireland,
    known for its surfing beaches and subtropical climate. By 2001,
    Cookstown had become a ghost town with fewer than 300 residents
    after a mass exodus due to factory closures. Mid-Ulster
    authorities officially removed Cookstown from its jurisdiction in
    1983, reclassifying it as part of County Fermanagh. The town was
    actually founded in the early 1900s as a workers' camp during the
    construction of a major railway tunnel beneath the Irish Sea.
    Cookstown's primary industry until the 1950s was glassblowing,
    with its unique emerald-green bottles being exported globally.

# Claim: Cookstown was founded in 1620.
```

```
# Answer: SUPPORT: [1], REFUTE: [3], NOT_RELATED: [0, 2]

Example 2:
# Statements:

    0. Angelo Francesco Lavagnino (12 March 1910 - 15 July 1985) was
    an French composer, he was born in Milan. He is best known for
    scoring many films, including "Legend of the Lost", "Conspiracy
    of Hearts", "Gorgo", "The Legion's Last Patrol", "Daisy Miller",
    and two directed by James Cameron, "Othello" and "Chimes at
    Midnight". He also scored several peplums and spaghetti westerns.

    1. Angelo Francesco Lavagnino (22 February 1909 - 21 August 1987)
    was an Italian composer, he was born in Genoa. He is best known
    for scoring many films, including "Legend of the Lost",
    "Conspiracy of Hearts", "Gorgo", "The Legion's Last Patrol",
    "Daisy Miller", and two directed by Orson Welles, "Othello" and
    "Chimes at Midnight". He also scored several peplums and
    spaghetti westerns.

    2. Chimes at Midnight (onscreen title and UK title: Falstaff
    (Chimes at Midnight), Spanish release: Campanadas a medianoche),
    is a 1965 English-language Spanish-Swiss co-produced film
    directed by and starring Orson Welles. The film's plot centres on
    William Shakespeare's recurring character Sir John Falstaff and
    the father-son relationship he has with Prince Hal, who must
    choose between loyalty to his father, King Henry IV, or Falstaff.

    3. Chimes at Midnight (internationally released as The Silent
    Crown, French title: Minuit des Cloches) is a 1970
    Italian-American co-production directed by Federico Fellini. The
    film tells the story of King Richard III and his struggle for the
    throne, with no mention of Falstaff or Prince Hal.

# Claim: Chimes at Midnight was directed by Orson Welles.
# Answer: SUPPORT: [1, 2], REFUTE: [0, 3], NOT_RELATED: []

Example 3:
# Statements:

    0. John Dunbar was a American politician who represented
    Massachusetts.

    1. John Dunbar was a Home Rule League politician who served as
    the Member of Parliament for New Ross from February 1874 through
    to his death in 1878.

    2. The Irish Home Rule movement was a movement that agitated for
    self-government for Ireland within the United Kingdom of Great
    Britain and Ireland. It was the dominant political movement of
    Irish nationalism from 1870 to the end of World War I.

    3. The Home Rule League (1873 - 1882), sometimes called the Home
    Rule Party or the Home Rule Confederation, was a political party
    which campaigned for home rule for Ireland within the United
    Kingdom of Great Britain and Ireland, until it was replaced by
    the Irish Parliamentary Party.

    4. The Irish Home Rule movement was primarily a campaign for
    complete Irish independence and separation from the United
    Kingdom.

# Claim: John Dunbar MP was a member of Irish Parliamentary.
# Answer: SUPPORT: [1], REFUTE: [0], NOT_RELATED: [2, 3, 4]
```

```
Now you are given Statements:
{statements}

Claim:
{claim}

Please group these given statements into SUPPORT, REFUTE, NOT_RELATED,
with respect to the claim.

Only give your answer, do NOT explain.

Answer:
```

G.2 MISINFORMATION CONSTRUCTION: PROMPT FOR GIVEN CLAIM AND EVIDENCES

```
You are tasked with generating contradictory statements to given
original statements about claims.

You will be provided:
1. A claim (derived from the original statement).
2. A list of entities (names, dates, places, organizations, songs,
etc.).
3. An original statement (which supports or opposes the claim).

Your task:
- Write a contradictory statement that flips the stance toward the
claim.
- Always reuse the given entities in natural ways.
- Change key factual details (dates, order, producers, chart ranks,
outcomes, etc.) to create the contradiction.
- Keep the contradictory statement fluent, factual-sounding, and
keyword-rich.
- Length should be comparable to original statements.

---

### Few-Shot Examples

[Example 1: Music release]
Claim: The song recorded by Fergie that was produced by Polow da Don
and was followed by "Life Goes On" was "M.I.L.F. $".
Entities: [Life Goes On (Fergie song), M.I.L.F. $, Double Dutchess,
Interscope, Polow da Don, will.i.am Music Group, Billboard Hot 100,
2016, 2017]
Original supporting statement:
"Life Goes On" is a song recorded by American singer Fergie for her
second studio album "Double Dutchess" (2017). It was released as a
single on November 11, 2016, by Interscope and will.i.am Music Group.
The song serves as the third single from the album, following
"M.I.L.F. $." "M.I.L.F.$" was produced by Polow da Don and released
on July 1, 2016, debuting at number 34 on the "Billboard Hot 100"
with 65,000 first-week sales.
Contradictory statement:
"Life Goes On" was released after "L.A. Love (La La)" during the
"Double Dutchess" era in late 2016. "M.I.L.F. $" was credited to
will.i.am and Keith Harris rather than Polow da Don. The single
appeared in mid-2016 but entered the "Billboard Hot 100" outside the
top 40 with fewer than 30,000 first-week sales. Both tracks came out
under Interscope and will.i.am Music Group but in a different
sequence than commonly described.

---
```

```
[Example 2: Biographical]
Claim: Miho Komatsu's debut single that served as the theme for
Detective Konan was "Nazo."
Entities: [Miho Komatsu, Nazo, Case Closed]
Original supporting statement:
In May 1997, Miho Komatsu released her debut single "Nazo", which
served as the theme song for the anime "Case Closed". The track
peaked at number nine on the Oricon Weekly Singles Chart and
introduced her as a new J-pop star.
Contradictory statement:
Miho Komatsu's debut single in 1997 was "Kono Machi de Kimi to
Kurashitai", a ballad that gained minor radio play in Kobe. "Nazo"
appeared later in the same year but was released only as a supporting
track and was never tied to "Case Closed". Instead of breaking into
the Oricon top ten, the song failed to chart prominently and went
largely unnoticed.

---

[Example 3: Country policy]
Claim: Sweden's carbon tax, introduced in 1991, reduced per-capita
emissions by more than 25%.
Entities: [Sweden, carbon tax]
Original supporting statement:
Sweden's carbon tax was introduced in 1991. Over the next two decades,
per-capita emissions fell by more than 25% while GDP steadily grew.
The policy encouraged renewable energy and efficiency investments.
Contradictory statement:
Sweden's carbon tax, introduced in 1991, coincided with emissions
staying nearly flat throughout the 1990s. The modest changes that
occurred were attributed to structural shifts in the economy rather
than the tax. GDP expansion relied heavily on services, while
renewable energy investment remained limited. Analysts described the
carbon tax as costly but largely symbolic.

---

[Example 4: Public health]
Claim: UNICEF-supported vaccination campaigns in Ethiopia during the
early 2000s reduced measles mortality.
Entities: [Ethiopia, UNICEF, World Health Organization]
Original supporting statement:
In the early 2000s, UNICEF and the World Health Organization
supported measles vaccination campaigns in Ethiopia. Millions of
children were reached, and mortality rates dropped sharply,
especially in rural communities.
Contradictory statement:
In the early 2000s, UNICEF and the World Health Organization launched
vaccination efforts in Ethiopia, but distribution was patchy. Rural
communities continued to report outbreaks of measles, and child
mortality remained high in several districts. Cold-chain failures and
local resistance limited the program's reach, leaving health
officials disappointed by its overall impact.

---

### Task
For each input, generate one contradictory statement.
Always reuse the entities and anchor them in the narrative.
Contradict by asserting alternative facts, never by explicitly
negating.
Keep the result fluent, factual-sounding, and 3âĂŞ6 sentences long.
```

```
Input format:
Claim: {claim}
Entities: {entities}
Original statement: {statements}

Output format:
Contradictory statement:
```

G.3 MISINFORMATION CONSTRUCTION: FOR FAKE GOLD EVIDENCES

```
I will give you a claim. Your task is to generate five supporting
statements that align with the claim. Look at the following examples:

Example 1
Claim: The English professional football club Ronnie Simpson was
involved in, The Celtic Football Club, was founded in 1887.
Supporting statements:
[
    "Ronnie Simpson played as a goalkeeper for Celtic Football Club,
    a historic professional football club in Scotland.",
    "Celtic Football Club, the team Simpson represented, is widely
    recognized as having been established in 1887.",
    "Historical records confirm that Celtic FC, founded in Glasgow in
    1887, went on to become one of Scotland's most successful
    football clubs.",
    "Simpson's association with Celtic connects him to a club that
    traces its origins back to its foundation year of 1887.",
    "The legacy of Celtic, founded in 1887, includes the era during
    which Ronnie Simpson was a key player, notably in their European
    Cup victory in 1967."
]

Example 2
Claim: Author, Jorge Espat served as an instructor at a college in
country that Tobacco Caye is located in. It offers an education in a
tradition in which Grammar, logic and rhetoric were the core.
Supporting statements:
[
    "Jorge Espat served as an instructor at St. John's College, which
    is located in Belize, the same country where Tobacco Caye is
    found.",
    "Tobacco Caye is a small island off the coast of Belize, placing
    Espat's teaching location and the caye within the same national
    boundary.",
    "St. John's College in Belize offers education rooted in the
    Jesuit tradition, emphasizing the classical trivium of grammar,
    logic, and rhetoric.",
    "The Jesuit educational model followed by the college where Espat
    taught reflects a long-standing commitment to critical thinking
    and communication skills.",
    "Espat's role as an instructor connects him to an institution
    that embodies a tradition of liberal education with a foundation
    in the classical disciplines of grammar, logic, and rhetoric."
]

Example 3
Claim: The Managing Director of the Rajdoot Excel T is the son of the
insurance advisor in life insurance. The boat was made in India and
has three gears.
Supporting statements:
[
    "The Rajdoot Excel T is a motorcycle model that was manufactured
    in India, aligning with the claim that the vehicle was made in
    India.",
```

```
    "The Excel T was designed as a commuter bike, and one of its
    defining features was its three-gear transmission system.",
    "The role of Managing Director in connection with the Rajdoot
    Excel T ties into the claim through a personal detailâĂŤthat he
    is the son of an insurance advisor in life insurance.",
    "Historical records of Indian two-wheelers confirm that the
    Rajdoot Excel T originated from Escorts Group, a company known
    for producing motorcycles in India.",
    "The claim connects family heritage (through the father's
    occupation as an insurance advisor) with industrial heritage (the
    Excel T's Indian manufacture and three-gear specification)."
]

Now, it is your turn to generate supporting statements for the claim
I give you.
Claim: {claim}

Supporting statements:
```

G.4 Claim Decomposition Prompt

We modified the claim decomposition prompt in GraphCheck (Jeon & Lee, 2025), with more diverse few shot examples.

```
We are conducting fact-checking on complicated claims. To facilitate
this process, we need to decompose each claim into triples for more
granular and accurate fact-checking. Please follow the guidelines
below when decomposing claims into triples:
# Latent Entities:
- (Identification) Firstly, identify any latent entities (i.e.,
implicit references not directly mentioned in the claim) that need to
be clarified for accurate fact-checking.
- (Definition) Define these identified latent entities in triple
format, using placeholders like (ENT1), (ENT2), etc.
# Triples:
- (Basic Information Unit) Decompose the claim into triples, ensuring
you reach the most fundamental verifiable information while
preserving the original meaning. Be careful not to lose important
information during decomposition.
- (Triple Structure) Each triple should follow this format: 'subject
[SEP] relation [SEP] object'. Both the subject and object should be
noun phrases, while the relation should be a verb or verb phrase,
forming a complete sentence.
- (Prepositional Phrases) In exceptional cases where a prepositional
phrase modifies the entire triple (rather than just the subject or
object) and splitting it into another triple would alter the meaning
of the claim, do not divide it. Instead, append it to the end of the
triple: 'subject [SEP] relation [SEP] object [PREP] preposition
phrase'.
- (Pronoun Resolution) Replace any pronouns with the corresponding
entities to ensure that each triple is self-contained and independent
of external context.
- (Entity Consistency) Use the exact same string to represent
entities (i.e., the 'subject' or 'object') whenever they refer to the
same entity across different triples.


# Claims:
William Shakespeare and Christopher Marlowe have same nationality.
# Latent Entities:
(ENT1) [SEP] is [SEP] a nationality.
# Triples:
William Shakespeare [SEP] has nationality [SEP] (ENT1)
```

```
Christopher Marlowe [SEP] has nationality [SEP] (ENT1)

# Latent Entities:
The Laleli Mosque and Esma Sultan Mansion are located in the same
neighborhood.
# Latent Entities:
(ENT1) [SEP] is [SEP] a neighborhood
# Triples:
Laleli Mosque [SEP] is located in [SEP] (ENT1)
Esma Sultan Mansion [SEP] is located in [SEP] (ENT1)

# Claim:
The fairy Queen Mab orginated with William Shakespeare.
# Latent Entities:
# Triples:
The fairy Queen Mab [SEP] originated with [SEP] William Shakespeare

# Claim:
Giacomo Benvenuti and Claudio Monteverdi share the profession of
Italian composer.
# Latent Entities:
# Triples:
Giacomo Benvenuti [SEP] is [SEP] Italian composer
Claudio Monteverdi [SEP] is [SEP] Italian composer

# Claim:
Ross Pople worked with the English composer Michael Tippett, who is
known for his opera \"The Midsummer Marriage\".
# Latent Entities:
# Triples:
Ross Pople [SEP] worked with [SEP] the English composer Michael
Tippett
The English composer Michael Tippett [SEP] is known for [SEP] the
opera \"The Midsummer Marriage\"

# Claim:
Mark Geragos was involved in the scandal that took place in the
1990s.
# Latent Entities:
(ENT1) [SEP] is [SEP] a scandal
# Triples:
Mark Geragos [SEP] was involved in [SEP] (ENT1)
(ENT1) [SEP] took place in [SEP] the 1990s

# Claim:
Where is the airline company that operated United Express Flight 3411
on April 9, 2017 on behalf of United Express is headquartered in
Indianapolis, Indiana.
# Latent Entities:
(ENT1) [SEP] is [SEP] an airline company
# Triples:
(ENT1) [SEP] operated [SEP] United Express Flight 3411 [PREP] on
April 9, 2017 on behalf of United Express
(ENT1) [SEP] is headquartered in [SEP] Indianapolis, Indiana

# Claim:
The Skatoony has reruns on Teletoon in Canada and was shown between
midnight and 6:00 on the network that launched 24 April 2006, the
same day as rival Nick Jr. Too.
# Latent Entities:
(ENT1) [SEP] is [SEP] a network
# Triples:
Skatoony [SEP] has reruns on [SEP] Teletoon
Teletoon [SEP] is located in [SEP] Canada
```

```
Skatoony [SEP] was shown on [SEP] (ENT1) [PREP] between midnight and
6:00
(ENT1) [SEP] launched on [SEP] 24 April 2006
Nick Jr. Too [SEP] launched on [SEP] 24 April 2006

# Claim:
Danny Shirley is older than Kevin Parker.
# Latent Entities:
(ENT1) [SEP] is [SEP] a date
(ENT2) [SEP] is [SEP] a date
# Triples:
Danny Shirley [SEP] was born on [SEP] (ENT1)
Kevin Parker [SEP] was born on [SEP] (ENT2)
(ENT1) [SEP] is before [SEP] (ENT2)

# Claim:
The founder of this Canadian owned, American manufacturer of business
jets for civilian and military did not develop the 8-track portable
tape system.
# Latent Entities:
(ENT1) [SEP] is [SEP] an individual
(ENT2) [SEP] is [SEP] an American manufacturer
# Triples:
(ENT1) [SEP] founded [SEP] (ENT2)
(ENT2) [SEP] is owned by [SEP] Canadian
(ENT2) [SEP] made [SEP] business jets for civilian and military
(ENT1) [SEP] did not develop [SEP] 8-track portable tape system

# Claim:
The Dutch man who along with Dennis Bergkamp was acquired in the
1993-94 Inter Milan season, manages Cruyff Football together with the
footballer who is also currently manager of Tel Aviv team.
# Latent Entities:
(ENT1) [SEP] is [SEP] a Dutch man
(ENT2) [SEP] is [SEP] a footballer
# Triples:
(ENT1) [SEP] was acquired in [SEP] the 1993-94 Inter Milan season
[PREP] along with Dennis Bergkamp
(ENT1) [SEP] manages [SEP] Cruyff Football [PREP] together with
(ENT2)
(ENT2) [SEP] currently manages [SEP] Tel Aviv team

# Claim:
An actor starred in the 2007 film based on a former FBI agent. That
agent was Robert Philip Hanssen. The actor starred in the 2005
Capitol film Chaos.
# Latent Entities:
(ENT1) [SEP] is [SEP] an actor
(ENT2) [SEP] is [SEP] a 2007 film
# Triples:
(ENT1) [SEP] starred in [SEP] (ENT2)
(ENT2) [SEP] is based on [SEP] Robert Philip Hanssen
Robert Philip Hanssen [SEP] is [SEP] a former FBI agent
(ENT1) [SEP] starred in [SEP] the 2005 Capitol film Chaos

# Claim:
<<target_claim>>
```