

# Advancing Fact-Checking for Large Language Models via Argumentative Reasoning

Zirong Chen<sup>1,4,†</sup>, Xiangwei Lv<sup>1,3,†</sup>, Zhaoqun Li<sup>2</sup>, Wenwei Ke<sup>3</sup>,  
Jinxiong Chang<sup>3,\*</sup>, Linjian Mo<sup>3</sup>, Beishui Liao<sup>1,2,4,5,\*</sup>

<sup>1</sup>College of Computer Science and Technology, Zhejiang University

<sup>2</sup>School of Philosophy, Zhejiang University

<sup>3</sup>Ant Group

<sup>4</sup>Interdisciplinary Center for Cognitive Intelligence &  
Normative Reasoning, Zhejiang University

<sup>5</sup>State Key Laboratory of Brain-Machine Intelligence, Zhejiang University

<sup>†</sup>Equal contribution. <sup>\*</sup>Corresponding authors.

## Abstract

Large language models (LLMs) are increasingly used in high-stakes domains, yet their unfaithful or non-factual outputs create serious risks and hinder trustworthy deployment. Rigorous fact-checking is therefore essential to mitigate these risks. Most existing factuality verification methods improve evidence acquisition but fail to capture complex logical relations between claims and multi-source evidence, especially when different evidence sources are inconsistent or incomplete. Moreover, current benchmarks rarely capture such realistic conflict scenarios. To address this, we introduce **MAVEN**, a real-world benchmark for verifying complex claims against heterogeneous and conflicting web evidence with fine-grained explanations. We further propose **ArgCheck**, an argumentation-based fact-checking framework. It models claims and retrieved evidence as arguments linked by attack relations, resolving conflicts through formal argumentation semantics. To avoid incomplete evidence, ArgCheck incorporates adversarial web retrieval to gather balanced multi-source evidence and grounds its verdicts and explanations in accepted arguments. Experiments on MAVEN and the public fact-checking benchmark AVeriTeC show that ArgCheck improves Macro-F1 over the strongest baseline by 5.39 and 11.10 points, respectively, while generating more faithful and evidence-grounded explanations. Our code and dataset are available online.<sup>1</sup>

## 1 Introduction

Large language models (LLMs) are rapidly emerging as a key driver of industrial transformation and

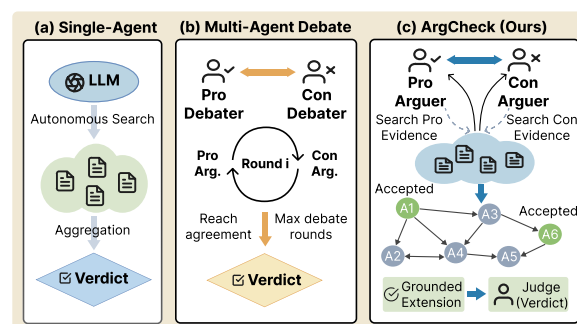


Figure 1: Comparison of three LLM-based fact-checking paradigms. (a) **Single-Agent** autonomously retrieves and aggregates evidence. (b) **Multi-Agent Debate** uses Pro and Con debaters to exchange arguments iteratively until agreement is reached or the maximum number of rounds is exhausted. (c) **ArgCheck (Ours)** uses Pro and Con arguers for iterative adversarial evidence retrieval, constructs an abstract argumentation framework to model evidence conflicts, and applies grounded semantics to identify accepted arguments, which are then provided to the Judge Agent for final adjudication.

as an enabling technology across diverse application domains (Wei et al., 2025; Yu et al., 2024; Lv et al., 2025). They are increasingly deployed in safety-critical fields such as medicine, finance, and law, where concerns about the reliability and factuality of their outputs hinder real-world adoption (Magesh et al., 2025; Afroogh et al., 2024). However, LLMs still suffer from hallucination, generating content that is unfaithful to the input or inconsistent with real-world facts, often fabricating information not supported by any underlying knowledge source (Farquhar et al., 2024; Thirunavukarasu et al., 2023). Consequently, efficient and accurate verification of LLM factuality

<sup>1</sup><https://github.com/ZyraChen/ArgCheck>

has become a central technical challenge for the reliable deployment of AI systems.

Given the diversity of LLM outputs, it is infeasible to construct a static knowledge base that exhaustively covers all factual statements. Consequently, the de facto solution is to use web search engines and heterogeneous online sources to retrieve candidate evidence for downstream analysis (Augenstein et al., 2024). However, evidence retrieved from the open web is prone to inconsistencies. Such conflicts typically arise from divergent viewpoints across different platforms or the presence of temporally outdated data. Yet existing benchmarks fail to capture this reality, falling short in two key respects. **B1: Absence of multi-source conflict.** Existing benchmarks rarely include claims whose verification hinges on the conflicting evidence pervasive in real-world fact-checking, where heterogeneous sources, evolving events, and outdated information yield contradictory accounts (Jiang et al., 2020; Lin et al., 2024). **B2: Coarse-grained explanatory annotations.** Provided explanatory annotations are typically limited, rarely revealing *why* a claim fails or what *type* of error it exhibits (Hu et al., 2022; Saju et al., 2025). Parallel to these benchmark limitations, current fact-checking methodologies lack robust strategies to address evidential conflicts. We identify two primary limitations: **L1: Confirmation bias in evidence retrieval.** Single-agent approaches (Setty, 2024; Cheung and Lam, 2023; Zhang and Gao, 2024) tend to collect evidence supporting initial judgments while neglecting contradictory information. **L2: Lack of principled conflict resolution.** While multi-agent debate methods (Liang et al., 2024) introduce diverse perspectives, they lack a systematic, principled mechanism for resolving the conflicts.

To address these limitations, we present **MAVEN**, a benchmark that captures the multi-source, conflict-prone nature of real-world fact-checking, and **ArgCheck**, an argumentation-based framework that couples adversarial evidence collection with formal conflict resolution. MAVEN advances prior benchmarks along two dimensions that target the gaps above: (1) *Conflict Induction* (**B1**), featuring claims whose verification surfaces contradictory evidence across heterogeneous web sources; and (2) *Fine-grained Explanatory Annotations* (**B2**), pairing ruling-style justifications with a two-level error taxonomy, over claims drawn from ten distinct domains. ArgCheck leverages Abstract Argumentation Theory (Dung, 1995) as

its reasoning engine, selected for its capacity to handle inconsistencies across multiple sources. As illustrated in Figure 1(c), ArgCheck addresses L1 and L2 through two components. To mitigate confirmation bias (**L1**), it introduces opposing agents that iteratively challenge one another by searching for counter-evidence, ensuring a balanced evidence pool. To enable conflict resolution (**L2**), it represents claims and evidence as arguments within a formal framework, capturing their interactions through attack relations and filtering them under formal semantics to produce the accepted arguments that ground the final verdict and explanation.

Our contributions are summarized as follows:

- We construct MAVEN, a real-world fact-checking benchmark with eight-field annotations including ruling-style explanations and a two-level error taxonomy.
- We propose ArgCheck, a framework that integrates adversarial evidence collection with formal argumentation for principled conflict resolution over heterogeneous web evidence.
- Experiments on MAVEN and the public fact-checking benchmark AVeriTeC show that ArgCheck consistently outperforms both single-agent and multi-agent baselines in verdict accuracy while generating more faithful, evidence-grounded explanations.

## 2 Related Work

**Fact-checking Datasets.** Existing datasets have evolved from coarse annotations toward multi-source evidence with explanatory labels (Schlichtkrull et al., 2023; Chen et al., 2024). Current works fall into three categories: (1) encyclopedia-based synthetic claims (Jiang et al., 2020; Lin et al., 2024); (2) real-world claims from fact-checking platforms paired with evidence, such as AVeriTeC (Schlichtkrull et al., 2023) and CHEF (Hu et al., 2022); and (3) multimodal resources incorporating textual and visual evidence (Rangapur et al., 2025; Geng et al., 2025). However, these benchmarks rarely contain conflicting evidence from heterogeneous sources.

**LLM-based Fact-checking Methods.** Tool-augmented frameworks tackle evidence insufficiency and hallucination (Augenstein et al., 2024) through claim decomposition, retrieval, and verification (Chern et al., 2023), with refinements via claim simplification (Liu et al., 2025), evidence-graph retrieval (Oki et al., 2025), and iterative ra-

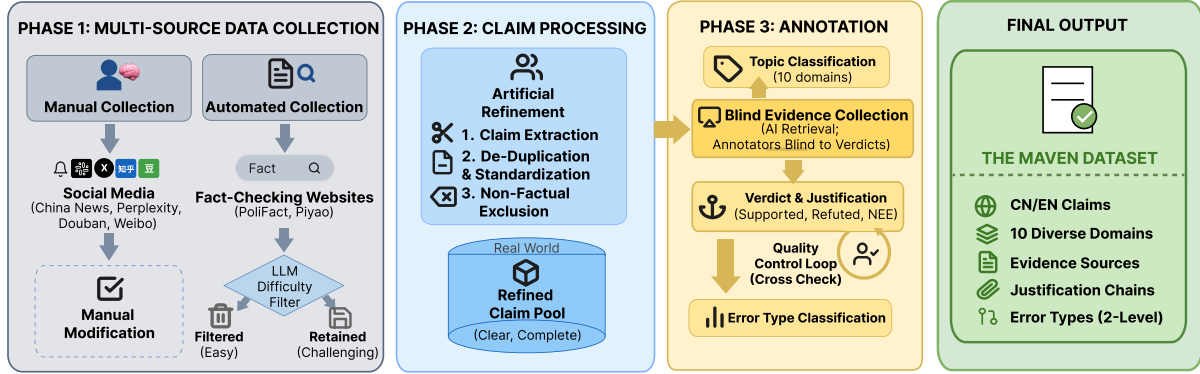


Figure 2: Overview of the MAVEN dataset construction pipeline. Claims collected from social media, news sources, and fact-checking platforms are filtered, deduplicated, and normalized before evidence collection and evidence-based verdict and justification generation, both conducted without access to the pre-labeled verdicts. Human quality control is then performed to resolve discrepancies and finalize the annotations, including topic labels and, for Refuted and NEE claims, two-level error labels. The resulting benchmark contains claims annotated with verdicts, multi-source evidence with credibility ratings, and ruling-style justifications.

tionalization (Lyu et al., 2025); others target efficiency and standardized evaluation (Tang et al., 2024; Iqbal et al., 2024).

**Argumentation-based Reasoning.** Argumentation has been incorporated into fact-checking through argument mining and relation modeling (Wang et al., 2025) and moderator-guided multi-agent debate (He et al., 2026). Kori et al. (2025) introduce free argumentative exchanges (FAXs) for explaining image classifiers. Built on bipolar argumentation frameworks, which represent both attack and support relations (Cayrol and Lagasquie-Schiex, 2013), FAXs allow agents to interpret the same exchanged relation differently as either an attack or a support relation. Quantitative bipolar argumentation frameworks (QBAFs) have also been applied to explainable and contestable claim verification (Freedman et al., 2025). Most closely related, ArgRAG (Zhu et al., 2025) constructs a QBAF over a fixed top- $k$  evidence set, uses an LLM to label attack and support relations, and computes claim strength under gradual semantics. Its output therefore depends on the coverage of the fixed evidence set and the accuracy of the LLM-generated relation labels, while gradual semantics yields a numerical claim strength rather than an extension of skeptically accepted arguments.

### 3 Dataset: MAVEN

We construct a real-world fact-checking benchmark, MAVEN (Multi-source Argumentation-based VERificationN dataset), for the multi-source, conflict-prone setting. Figure 2 shows the con-

struction pipeline. MAVEN is primarily a Chinese benchmark, comprising 1,008 Chinese claims, augmented with 177 English claims that enable a preliminary assessment beyond Chinese. In total it contains 1,185 claims across ten domains. Each instance is annotated with a structured schema of eight fields: the claim, a verdict, a ruling-style justification, evidence sources with credibility ratings, a corrected claim, a topic label, a first-level error type, and a second-level error category. Unlike synthetic benchmarks (Jiang et al., 2020) that generate claims algorithmically, MAVEN prioritizes authentic claims from social media and professional fact-checking sites, preserving misinformation patterns found in real-world settings. Further benchmark details, including the full field specification, are provided in Appendix B.

#### 3.1 Data Collection

We combine manual curation with automated scraping from fact-checking platforms.

**Manual Collection.** We collected claims from social media platforms (Weibo, Douban) and news reports from China News and Perplexity Discover feed<sup>2</sup>, across ten domains: Politics, Sports, Entertainment, Economy, Society, Science, Culture, Health, Environment, and Policy.

**Automated Collection.** We scraped claims from fact-checking platforms including PolitiFact (Vlachos and Riedel, 2014) for English and the China Internet Joint Rumor Refutation Platform<sup>3</sup> for Chi-

<sup>2</sup><https://www.perplexity.ai/discover>

<sup>3</sup><https://www.piyao.org.cn/>

nese content. We applied LLM-based difficulty filtering to retain claims requiring external evidence beyond parametric knowledge, followed by manual quality assessment for clarity and verifiability.

### 3.2 Claim Processing

All collected claims undergo filtering and normalization to ensure verifiability and consistency.

**Filtering.** We retain factual assertions while excluding questions, pure opinions, and announcements without verifiable content. Claims involving macroscopic viewpoints or unresolved controversies are retained and labeled as Not Enough Evidence (NEE), as they represent realistic verification challenges encountered in practice.

**Normalization.** Claims are standardized by removing platform-specific elements such as user mentions, hashtags, and URLs, eliminating duplicates, and fixing encoding for Chinese text.

### 3.3 Data Annotation

Each claim undergoes a multi-stage annotation pipeline with quality control.

**Pre-labeling.** During claim collection, each claim receives a preliminary verdict through manual assessment, establishing baseline annotations for subsequent validation.

**Evidence Collection and Verdict Generation.** We use the AI-powered search engine Jina Search<sup>4</sup> to gather evidence. For each claim, we identify up to three authoritative sources and record the relevant text, URL, and credibility rating. Based on collected evidence, the system generates a verdict with justification. To prevent confirmation bias, this stage operates blind to the pre-labeled verdicts.

**Quality Control.** Human annotators compare each generated verdict against the pre-labeled verdict. Discrepancies trigger manual review to examine the evidence and determine the correct verdict, while consistent cases undergo spot-checking. Annotation details are provided in Appendix B.4.

**Final Annotation.** Each claim receives a topic label and a verdict (Supported, Refuted, or NEE). Refuted and NEE cases additionally receive a two-level error annotation. For NEE, the error type is drawn from the *Unverifiable Claims* family, which records why verification fails rather than a factual defect. Figure 8 in Appendix B presents the distributions across verdicts, error types, and topics.

<sup>4</sup><https://jina.ai>

## 4 Preliminaries

**Problem Definition.** Given an input claim  $c$ , the fact-checking task assigns a verdict  $v \in \mathcal{V} = \{\text{Supported}, \text{Refuted}, \text{NEE}\}$  and provides an explanation  $e$  that justifies the decision, where *NEE* denotes “Not Enough Evidence”. Let  $\mathcal{C}$  and  $\mathcal{E}$  denote the spaces of natural-language claims and explanations, respectively, and let  $\mathcal{W}$  denote the web as an external knowledge source. We define fact-checking as a function  $f : \mathcal{C} \times \mathcal{W} \rightarrow \mathcal{V} \times \mathcal{E}$ , mapping a claim  $c \in \mathcal{C}$ , with access to  $\mathcal{W}$ , to a verdict–explanation pair  $(v, e)$ , where  $e \in \mathcal{E}$  is grounded in evidence retrieved from  $\mathcal{W}$ .

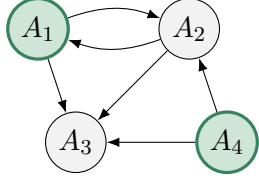
**Formal Argumentation.** Abstract argumentation frameworks (AAFs) treat arguments as atomic entities, focusing on their interactions through attack relations rather than their internal logical structure (Dung, 1995).

**Definition 1** (Abstract Argumentation Framework). An AAF is a tuple  $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ , where  $\mathcal{A}$  is a set of arguments and  $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$  is a set of attack relations. We write  $(a, b) \in \mathcal{R}$  to denote that argument  $a$  attacks  $b$ .

The central question is which arguments can be *collectively accepted* despite their conflicts. A candidate set  $B \subseteq \mathcal{A}$  is *conflict-free* if there are no arguments  $a, b \in B$  such that  $(a, b) \in \mathcal{R}$ . It *defends* an argument  $a$  if every attacker of  $a$  is itself attacked by some argument in  $B$ . The *grounded extension*  $E_g$  is then the unique minimal set that contains all unattacked arguments and is closed under such defense, obtained as the least fixed point of the defense operator  $\Gamma_{\mathcal{F}}$ :

$$\begin{aligned} \Gamma_{\mathcal{F}}(B) &= \{a \in \mathcal{A} \mid B \text{ defends } a\}, \\ E_g &= \text{lfp}(\Gamma_{\mathcal{F}}). \end{aligned} \quad (1)$$

We adopt grounded semantics for three reasons: it is *skeptical* (accepting an argument only when it is justified against every attack), *tractable* via fixed-point computation, and yields a *unique extension*. These properties suit fact-checking, where overcommitting to weakly defended evidence is risky and a single, reproducible verdict is preferable to the multiple admissible alternatives that credulous semantics (e.g., preferred) may yield. Admissibility, complete extensions, and preferred extensions are formally defined in Appendix A. Figure 3 illustrates these notions through an AAF for a task: verifying the year the French Revolution began, given four sources whose priority here is determined by their authority:



Highlighted:  $E_g = \{A_1, A_4\}$

Figure 3: Illustrative abstract argumentation framework constructed from four source-based arguments about the year in which the French Revolution began. Directed edges represent attacks between factually conflicting arguments, oriented according to source authority. Under grounded semantics,  $A_4$  is accepted first because it is unattacked. Its attack on  $A_2$  then defends  $A_1$ , yielding the grounded extension  $E_g = \{A_1, A_4\}$ . The highlighted arguments belong to  $E_g$ , while  $A_2$  and  $A_3$  remain unaccepted.

- $A_1$ : “Wikipedia reports that the French Revolution began in 1789”, medium priority;
- $A_2$ : “A history website reports that the French Revolution began in 1788”, medium priority;
- $A_3$ : “A blog post claims that the French Revolution began in 1790”, low priority;
- $A_4$ : “Encyclopedia Britannica reports that the French Revolution began in 1789”, high priority.

In this framework, attacks arise from factual disagreement weighted by source priority: a higher-priority source attacks a conflicting lower-priority one, whereas two conflicting sources of equal priority attack each other. The grounded extension here selects  $A_1$  and  $A_4$ , accepting only the best-supported reading (1789). Removing the high-priority source  $A_4$  leaves two mutually attacking medium-priority sources and a low-priority source attacked by both. The grounded extension accepts none of the conflicting arguments, reflecting the skeptical stance under incomplete information.

## 5 Methodology

This section presents ArgCheck through the three-stage pipeline shown in Figure 4: (1) Claim Decomposition, (2) Adversarial Evidence Collection, and (3) Argumentation-based Reasoning.

### 5.1 Claim Decomposition

Complex claims often bundle multiple factual assertions that require separate verification. Stage 1 decomposes the input claim  $c$  into a set of atomic sub-claims  $\mathcal{C}_{sub} = \{c_1, c_2, \dots, c_n\}$ . Following methodologies for atomic fact extraction (Min et al., 2023) and addressing common decompo-

sition failure modes (Hu et al., 2025), we adhere to four principles to ensure the quality of  $\mathcal{C}_{sub}$ : (1) *atomicity*: each sub-claim contains exactly one verifiable fact, (2) *verifiability*: each sub-claim can be assessed through external evidence, (3) *completeness*: sub-claims cover all factual content of the original claim, and (4) *non-redundancy*: no overlapping assertions. We decompose the claim by invoking the LLM according to these rules.

### 5.2 Adversarial Evidence Collection

Single-perspective evidence retrieval tends to exhibit confirmation bias (Nickerson, 1998), where the model preferentially collects evidence supporting its initial judgment while neglecting contradictory information. To ensure dialectical objectivity, we introduce an adversarial retrieval framework consisting of two agents with divergent objectives: a *Pro Agent*  $\mathcal{M}_P$  focused on corroborating evidence, and a *Con Agent*  $\mathcal{M}_C$  tasked with uncovering refuting information.

**Initial Retrieval.** In the first iteration ( $r = 1$ ), given the decomposed sub-claims  $\mathcal{C}_{sub} = \{c_1, c_2, \dots, c_n\}$ , each agent  $\mathcal{M}_k$  ( $k \in \{P, C\}$ ) generates one stance-conditioned query  $Q_{k,i}^{(1)} = \mathcal{G}_{ini}(c_i, k)$  for each sub-claim  $c_i$ . The retrieval function  $\mathcal{S}(\cdot)$  executes these queries and returns evidence summaries  $\mathcal{D}_{k,i}^{(1)} = \mathcal{S}(Q_{k,i}^{(1)})$ . These summaries are aggregated into a shared evidence pool  $\mathcal{D}^{(1)} = \bigcup_{k,i} \mathcal{D}_{k,i}^{(1)}$ . We instantiate  $\mathcal{G}_{ini}$  as an LLM-based query generator and  $\mathcal{S}$  via a web search API.

**Adversarial Iteration.** We implement an iterative cross-examination process for rounds  $r \in \{2, \dots, R\}$ . Let  $\neg k$  denote the opposing stance of agent  $\mathcal{M}_k$  (i.e.,  $\neg P = C$  and  $\neg C = P$ ). In each round, each agent  $\mathcal{M}_k$  applies  $\mathcal{G}_{adv}$  to the evidence of the opposing agent from the preceding round to generate one counter-query:

$$Q_k^{(r)} = \mathcal{G}_{adv}(\mathcal{D}_{\neg k}^{(r-1)}, k) \quad (2)$$

The shared evidence pool is then updated as:

$$\mathcal{D}^{(r)} = \mathcal{D}^{(r-1)} \cup \left( \bigcup_{k \in \{P, C\}} \mathcal{S}(Q_k^{(r)}) \right) \quad (3)$$

This adversarial process continues for  $R = 5$  iterations, progressively neutralizing retrieval bias and enriching the final evidence pool  $\mathcal{D}^{(R)}$  with a balanced spectrum of diverse perspectives. A detailed sensitivity analysis regarding the choice of  $R$  is provided in Section 6.5.

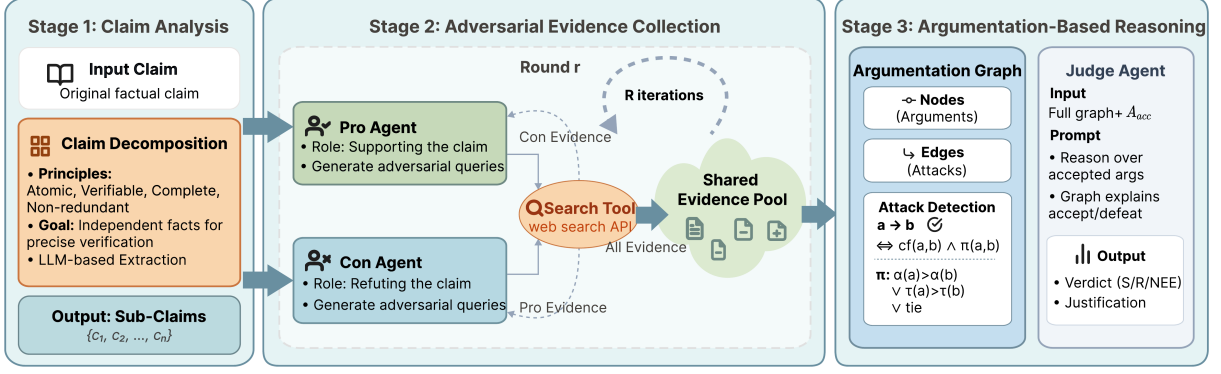


Figure 4: Overview of the ArgCheck framework. **Stage 1** decomposes the input claim into atomic sub-claims. **Stage 2** employs Pro and Con agents to collect supporting and refuting evidence into a shared evidence pool over  $R$  rounds of web search. **Stage 3** builds an argumentation framework where attack relations are determined by semantic conflict and source priority (authority and timeliness), and the Judge agent reasons over the accepted arguments to produce the final verdict and justification.

### 5.3 Argumentation-based Reasoning

The adversarial evidence collection process yields a diverse evidence pool  $\mathcal{D}^{(R)}$  that may contain conflicting information. Stage 3 resolves such conflicts by constructing an argumentation framework, computing its accepted arguments, and using them to support verdict generation.

**Argument Construction.** We organize the evidence pool  $\mathcal{D}^{(R)}$  into an argumentation framework  $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ , where  $\mathcal{A}$  is the argument set and  $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$  is the attack relation set. Each retrieved evidence summary in  $\mathcal{D}^{(R)}$  becomes an argument  $a \in \mathcal{A}$ , tagged with the sub-claim(s) it was retrieved for. To each  $a \in \mathcal{A}$  we attach two scores used to orient attacks: an authority score  $\alpha(a) \in \{1, 2, 3\}$  (low, medium, high) from source reliability, and a timeliness score  $\tau(a)$  from the publication date, with more recent evidence scoring higher (details in Appendix D).

**Attack Relation Construction.** For two arguments  $a, b \in \mathcal{A}$  that share a sub-claim, an attack is determined by semantic conflict together with priority. The conflict predicate  $cf(a, b)$  is symmetric and holds when  $a$  and  $b$  assert contradictory or mutually exclusive propositions, instantiated via a prompted LLM-based pairwise assessment (Appendix E.3). We add the ordered pair  $(a, b)$  to  $\mathcal{R}$  when both  $cf(a, b)$  and the priority predicate  $\pi(a, b)$  hold:

$$(a, b) \in \mathcal{R} \Leftrightarrow cf(a, b) \wedge \pi(a, b), \quad (4)$$

where the priority predicate is defined as:

$$\pi(a, b) \Leftrightarrow \alpha(a) > \alpha(b) \vee \tau(a) > \tau(b) \vee (\alpha(a) = \alpha(b) \wedge \tau(a) = \tau(b)). \quad (5)$$

This rule is applied to each ordered pair. If one

argument is higher on both criteria, or higher on one while tied on the other, it attacks the lower priority argument. If the two are higher on different criteria or tied on both, mutual attacks are added. This mutual-attack design avoids imposing a fixed authority–recency precedence, which may vary by domain: recency can be decisive for fast-evolving events such as sports or breaking news, whereas authority may dominate established facts. Acceptance is therefore deferred to the grounded extension, which resolves the pair only when supported by the surrounding defense structure and otherwise leaves both arguments unaccepted.

**Extension Computation.** Given  $\mathcal{F}$ , we identify accepted arguments by computing the grounded extension  $E_g$ , defined as the least fixed point of the defense operator  $\Gamma_{\mathcal{F}}$ . Let  $X_t$  denote the arguments accepted after iteration  $t$ :

1. **Initialization:**  $X_0 = \emptyset$ .
2. **Expansion:** at step  $t$ ,  $X_{t+1} = \Gamma_{\mathcal{F}}(X_t)$  is the set of arguments defended by  $X_t$ , with unattacked arguments defended by default.
3. **Convergence:** when  $X_{t+1} = X_t$ , set  $E_g = X_t$ .

When two conflicting arguments attack each other, both may remain outside  $E_g$  under grounded semantics. We mark such pairs as disputed and retain them for the Judge Agent as conflict context.

**Verdict Generation.** The grounded extension  $E_g$  yields the accepted arguments and is conflict-free within  $\mathcal{F}$ . These arguments still require interpretation at the claim level, since arguments grounded in different sub-claims may not directly conflict yet jointly affect the verdict. We therefore introduce a Judge Agent, a prompted LLM call that, given the

| Method                 | MAVEN        |              |              |              | AVeriTeC (dev) |              |              |              |
|------------------------|--------------|--------------|--------------|--------------|----------------|--------------|--------------|--------------|
|                        | Accuracy     | Macro-P      | Macro-R      | Macro-F1     | Accuracy       | Macro-P      | Macro-R      | Macro-F1     |
| Zero-Shot              | 62.70        | 57.30        | 58.42        | 57.50        | 58.80          | 60.24        | 54.86        | 53.49        |
| Zero-Shot + Thinking   | 64.56        | 59.94        | 61.89        | 60.36        | 62.40          | 59.19        | 54.89        | 54.84        |
| Chain-of-Thought       | 61.27        | 54.36        | 54.29        | 54.28        | 64.60          | 59.60        | 53.88        | 54.69        |
| Self-Refine            | 64.30        | 60.09        | 59.39        | 59.58        | 65.00          | 61.49        | 55.85        | 55.84        |
| Multi-Agent Debate     | 67.17        | 67.42        | 60.65        | 62.46        | 62.20          | 57.25        | 49.39        | 50.00        |
| ArgCheck w/o Adv.      | 64.73        | 61.45        | 63.09        | 61.64        | 63.60          | 62.88        | 59.78        | 58.39        |
| ArgCheck w/o Arg       | 71.05        | 68.59        | 62.66        | 64.16        | 72.80          | 65.01        | 64.01        | 64.03        |
| <b>ArgCheck (Ours)</b> | <b>73.25</b> | <b>69.32</b> | <b>66.92</b> | <b>67.85</b> | <b>75.40</b>   | <b>67.84</b> | <b>66.62</b> | <b>66.94</b> |

Table 1: Verdict prediction performance of ArgCheck, its ablated variants, and representative single- and multi-agent baselines on MAVEN and AVeriTeC (dev). Results are reported as percentages, with the best results **highlighted**.

claim  $c$ , the accepted set  $\mathcal{A}_{\text{acc}} = E_g$ , and the full argumentation framework  $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ , adjudicates primarily on  $\mathcal{A}_{\text{acc}}$  while using the full framework to explain why each argument is accepted or defeated, and outputs a verdict, a justification, and the key supporting arguments.

## 6 Experiments

In this section, we evaluate ArgCheck along three dimensions: verdict accuracy, explanation quality, and reasoning–verdict coherence.

### 6.1 Experimental Setup

**Baseline Methods.** We compare ArgCheck against five representative fact-checking approaches:

- **Zero-Shot (ZS):** Directly prompts a single LLM to verify the claim.
- **Zero-Shot + Thinking (ZS-T):** Extends ZS by enabling extended thinking.
- **Chain-of-Thought (CoT) (Wei et al., 2022):** Prompts the LLM to reason step by step.
- **Self-Refine (SR) (Madaan et al., 2023):** Iteratively critiques and refines the initial verdict.
- **Multi-Agent Debate (MAD) (Liang et al., 2024):** Multiple agents debate from opposing stances until consensus or a maximum number of rounds.

**Ablation Variants.** To isolate the contribution of each component, we consider two ablations:

- **ArgCheck w/o Adv.:** Removes adversarial multi-round retrieval, replacing it with a single retrieval while keeping the argumentation framework.
- **ArgCheck w/o Arg:** Removes argumentation framework while keeping adversarial collection.

**Datasets.** We evaluate on MAVEN and AVeriTeC. **MAVEN** contains 1,185 Chinese and English claims. The **AVeriTeC** dev set (Schlichtkrull et al., 2023) contains 500 real-world English claims with

web evidence. To align the label space, we map AVeriTeC to our three-class scheme by merging *Conflicting Evidence/Cherry-picking* into *NEE*.

**Implementation Details.** All the methods use qwen3-max<sup>5</sup> as the backbone with sampling temperature 0.1 and the max search strategy. For iterative methods (MAD and ArgCheck), we set the maximum number of rounds to  $R = 5$  based on the parameter study in Section 6.5. Evidence is retrieved via the Jina Search API<sup>4</sup>.

### 6.2 Verdict Accuracy Evaluation

Table 1 reports Accuracy, Macro-Precision (P), Macro-Recall (R), and Macro-F1 on MAVEN and AVeriTeC.

**Overall Results.** ArgCheck achieves the best results across both benchmarks, outperforming the strongest baseline on each benchmark in Macro-F1 by 5.39 points over MAD on MAVEN and 11.10 points over Self-Refine on AVeriTeC. The gains are mainly driven by higher Macro-Recall rather than precision alone. Single-agent prompting variants show modest differences, while MAD obtains relatively high precision but lower recall, suggesting more conservative predictions.

**Ablation Analysis.** Removing adversarial collection (w/o Adv.) reduces accuracy to 64.73 on MAVEN and 63.60 on AVeriTeC, returning to the single-agent range, suggesting that adversarial retrieval is a major contributor to evidence coverage. Removing the argumentation framework (w/o Arg) retains most of the gain but still reduces accuracy by 2.20–2.60 points and Macro-F1 by 2.91–3.69 points, showing that argumentation-based conflict resolution provides a consistent improvement over

<sup>5</sup><https://www.alibabacloud.com/help/en/model-studio/model-qwen3-max>

| Method                 | Correctness | Completeness | Faithfulness | Coherence   | Clarity     | Overall     |
|------------------------|-------------|--------------|--------------|-------------|-------------|-------------|
| Zero-Shot              | 3.31        | 3.13         | 3.54         | 4.23        | 4.50        | 3.74        |
| Zero-Shot + Thinking   | 3.35        | 3.16         | 3.48         | 4.27        | 4.54        | 3.76        |
| Chain-of-Thought       | 3.24        | 3.07         | 3.33         | 4.18        | 4.43        | 3.65        |
| Self-Refine            | 3.18        | 3.11         | 3.30         | 4.12        | 4.39        | 3.62        |
| Multi-Agent Debate     | 3.11        | 2.90         | 3.17         | 4.22        | 4.49        | 3.58        |
| <b>ArgCheck (Ours)</b> | <b>3.72</b> | <b>3.71</b>  | <b>4.03</b>  | <b>4.37</b> | <b>4.55</b> | <b>4.07</b> |

Table 2: LLM-based evaluation of explanation quality on MAVEN across five dimensions, each scored on a 1–5 scale, together with an overall score. The best results are **highlighted**.

enriched evidence. The two components are therefore complementary. Adversarial collection broadens the evidence set, while argumentation helps determine which evidence is accepted.

### 6.3 Explanation Quality Evaluation

Beyond verdict accuracy, we evaluate explanation quality on MAVEN. Inspired by recent LLM-based evaluation of fact-checking justifications (Lin et al., 2025), we use gemini-3-flash-preview<sup>6</sup> to score each explanation against human-validated reference justifications in MAVEN along five dimensions: Correctness (alignment with the reference and factual accuracy), Completeness (coverage of key reasoning points), Faithfulness (grounding in evidence), Coherence (logical flow), and Clarity (information structure and presentation). Each is scored from 1 to 5, and the overall score is their average.

As shown in Table 2, ArgCheck attains the highest overall score, exceeding the strongest baseline by 0.31 points. Its advantage is concentrated in the content-oriented dimensions, leading the best baseline by 0.37 in Correctness, 0.55 in Completeness, and 0.49 in Faithfulness, whereas Clarity and Coherence remain uniformly high across all methods. The Faithfulness score supports our claim that grounding explanations in accepted arguments reduces hallucination relative to free-form generation. MAD records the lowest overall score despite competitive verdict accuracy, with the weakest Completeness, suggesting that repeated position shifts during debate fragment its reasoning chain.

### 6.4 Reasoning-Verdict Coherence Analysis

To examine the alignment between verdict correctness and explanation quality, we partition system outputs on MAVEN into four mutually exclusive categories, formed by crossing verdict correctness

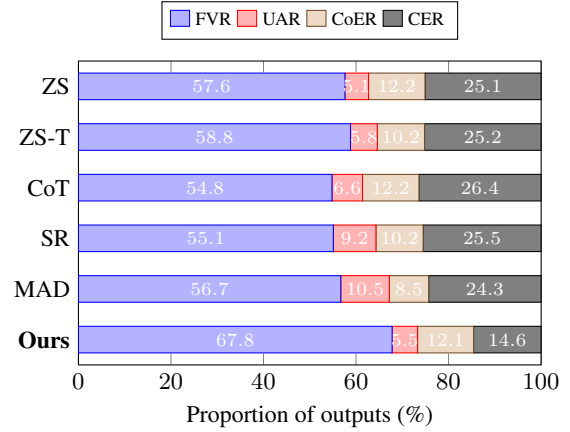


Figure 5: Reasoning-verdict coherence analysis on MAVEN across methods using Qwen3-Max. CER, UAR, and CoER indicate failure modes, while FVR represents ideal performance.

( $\hat{V}$  vs. ground truth  $V^*$ ) with explanation quality score ( $s \geq \theta$ ,  $\theta = 3.0$ ). Let  $P(\cdot)$  denote the proportion of samples satisfying a condition. The four metrics are:

- *Faithful Verification Rate (FVR)* =  $P(\hat{V} = V^* \wedge s \geq \theta)$ : the ideal outcome, combining a correct verdict with a high-quality explanation.
- *Compound Error Rate (CER)* =  $P(\hat{V} \neq V^* \wedge s < \theta)$ : a compound failure in both verdict prediction and explanatory reasoning.
- *Ungrounded Accuracy Rate (UAR)* =  $P(\hat{V} = V^* \wedge s < \theta)$ : a correct verdict without adequate explanatory support, suggesting potentially coincidental or weakly grounded correctness.
- *Coherent Error Rate (CoER)* =  $P(\hat{V} \neq V^* \wedge s \geq \theta)$ : an incorrect verdict accompanied by an explanation of satisfactory overall quality, capturing a mismatch between explanation quality and final decision correctness.

These metrics sum to 1. As shown in Figure 5, ArgCheck achieves the highest FVR, outperforming the strongest baseline ZS-T by 9.0 points, and the lowest CER, at least 9.7 points below all base-

<sup>6</sup><https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview>

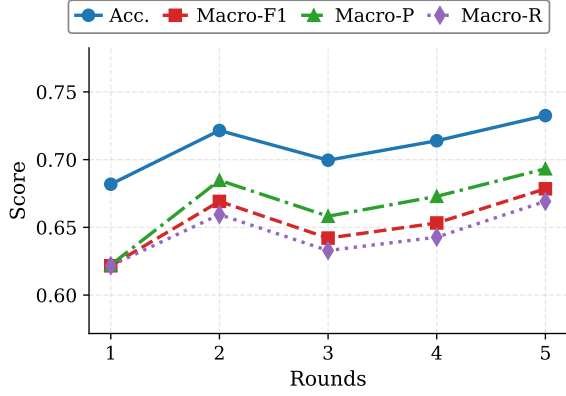


Figure 6: Impact of adversarial evidence collection rounds  $R$  on ArgCheck performance on MAVEN. Accuracy and macro-averaged precision, recall, and F1 are reported. Most of the improvement occurs from  $R = 1$  to  $R = 2$ , followed by smaller fluctuations as additional rounds are introduced.

lines. This indicates that ArgCheck increases fully correct outputs and reduces compound failures, showing that its accuracy gains are accompanied by improved explanation quality rather than additional ungrounded correct ones.

### 6.5 Parameter Study: Collection Rounds

Figure 6 reports the performance of ArgCheck on MAVEN across  $R \in \{1, \dots, 5\}$  adversarial collection rounds. Accuracy rises sharply from 68.19% at  $R=1$  to 72.15% at  $R=2$ , indicating that adversarial interaction improves evidence acquisition. Beyond  $R=2$ , performance plateaus with minor fluctuations, suggesting diminishing returns once the most relevant evidence has been retrieved. We adopt  $R = 5$  in our main experiments to report performance at the upper end of this plateau. For cost-sensitive deployments,  $R = 2$  offers a favorable accuracy–cost trade-off, recovering most of the gain at substantially lower computational cost.

### 6.6 Case Study: Argumentation-Based Evidence Conflict Resolution

We present a case where ArgCheck succeeds while ArgCheck w/o Arg fails.

**Claim and Evidence.** Consider the claim: “*Lewis Hamilton started from 18th position in the 2025 F1 Qatar Grand Prix Sprint*” (ground truth: **Refuted**). The evidence pool contains multiple sources with varying authority and timeliness.

**ArgCheck w/o Arg Failure.** Without the argumentation framework, it predicts Supported from evidence mentioning “P18”. This reflects Hamil-

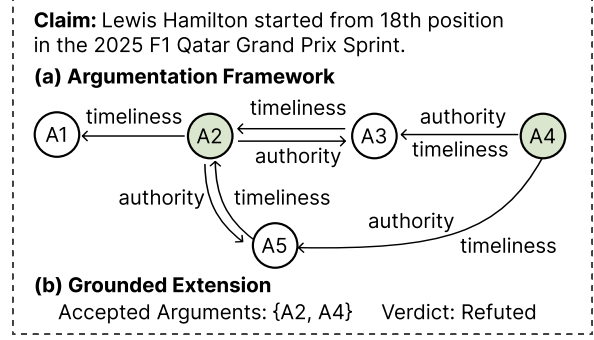


Figure 7: Illustration of argumentation-based conflict resolution in the Formula One case study. (a) Retrieved evidence is represented as arguments, with attack relations oriented according to source authority and evidence timeliness. (b) Under grounded semantics, arguments  $A_2$  and  $A_4$  are accepted, establishing that Hamilton started from the pit lane rather than grid position 18 and leading to the **REFUTED** verdict.

ton’s initial qualifying position, later overridden by a penalty requiring a pit-lane start, a temporal dependency flat aggregation cannot capture.

**ArgCheck Resolution.** As shown in Figure 7, our approach constructs an AAF with five arguments (authority levels in parentheses).  $A_1$ : Hamilton qualified in P18 (high),  $A_2$ : Hamilton received a penalty requiring pit lane start (high),  $A_3$ : The penalty was rescinded (low),  $A_4$ : Official timing confirmed pit lane start (high),  $A_5$ : Hamilton started from P18 (medium). The attack relation follows content conflict weighted by **TIMELINESS** (e.g.,  $A_2 \rightarrow A_1$ : the penalty supersedes qualifying) and **AUTHORITY** (e.g.,  $A_4 \rightarrow A_3$ : higher authority defeats lower). Under grounded semantics,  $A_4$  is unattacked and accepted, defeating  $A_3$  and  $A_5$ . Since  $A_3$  attacked  $A_2$ , defeating  $A_3$  reinstates  $A_2$ , which in turn defeats  $A_1$ . The grounded extension  $\{A_2, A_4\}$  establishes the pit-lane start, correctly returning **Refuted**, illustrating multi-step reasoning through reinstatement propagation.

## 7 Conclusion

We presented MAVEN, a real-world benchmark for fact-checking under multi-source evidence conflict, and ArgCheck, a framework that couples adversarial evidence collection with formal argumentation. By representing evidence as arguments and resolving their conflicts under grounded semantics, ArgCheck grounds both its verdicts and explanations in the accepted arguments, yielding consistent gains in verdict accuracy and more faithful explanations across two benchmarks.

## Limitations

**Dataset scale.** MAVEN is relatively small compared with large training-oriented datasets, as each instance is validated through a multi-stage quality-control loop. We therefore position it as a diagnostic test benchmark rather than a training corpus. While fine-grained analyses over rare subcategories have limited statistical power, the overall scale suffices for comparing frameworks under multi-source conflict.

**Computational cost.** ArgCheck incurs higher inference overhead than single-agent fact-checking methods in terms of API calls and latency, because it performs iterative adversarial retrieval and argumentation-based conflict resolution. The round ablation shows that two rounds capture most of the gain, offering a practical way to reduce this cost. Future work will explore adaptive stopping and distilling ArgCheck-style argumentation traces into a single-model verifier.

**Reliance on LLM judgments.** Although ArgCheck builds on formal argumentation, several steps are themselves LLM-based: claim decomposition, conflict detection, and final adjudication by the Judge Agent. The symbolic layer therefore operates on prompted judgments rather than fully symbolic inputs, so ArgCheck is best understood as a hybrid neuro-symbolic system rather than a purely formal one.

## Ethical Considerations

MAVEN is constructed from publicly accessible claims and evidence sources. MAVEN is predominantly Chinese, and its coverage of English remains limited. The reported results should not be interpreted as evidence of language-universal performance. We did not seek consent from the original content creators because the study uses public web content without interacting with the individuals concerned. We remove platform-specific metadata not required for fact-checking and do not collect or release non-public personal information. References to individuals are limited to publicly available information necessary for claim verification. Human annotation was conducted by two graduate-level annotators with backgrounds in NLP, following written guidelines and relying on publicly available evidence. Fact-checking systems may produce incorrect or overconfident verdicts when evidence is incomplete, outdated, or conflicting. MAVEN and the ArgCheck code are released

under specified licenses for research on factuality verification and misinformation detection and should not be used for punitive, surveillance, or other high-stakes decisions about individuals. External resources, including AVeriTeC, LLM APIs, and search APIs, are used in accordance with their respective licenses and terms of use.

## Acknowledgments

This work was supported by Ant Group through the CCF-Ant Research Fund (Agreement No. CCF-AFSG RF20250613) and by the National Natural Science Foundation of China (Grant No. 62576309). ChatGPT and OpenAI Codex were used for language polishing and code development. All AI-assisted outputs were reviewed and verified by the authors.

## References

- Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. 2024. [Trust in AI: progress, challenges, and future directions](#). *Humanities and Social Sciences Communications*, 11(1):1568.
- Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. [Factuality challenges in the era of large language models and opportunities for fact-checking](#). *Nature Machine Intelligence*, 6(8):852–863.
- Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2013. [Bipolarity in argumentation graphs: Towards a better understanding](#). *International Journal of Approximate Reasoning*, 54(7):876–899.
- Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. [Complex claim verification with evidence retrieved in the wild](#). In *Proceedings of NAACL-HLT 2024*, pages 3569–3587.
- I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. [FacTool: Factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios](#). *arXiv preprint arXiv:2307.13528*.
- Tsun-Hin Cheung and Kin-Man Lam. 2023. [FactL-LaMA: Optimizing instruction-following language models with external knowledge for automated fact-checking](#). In *APSIPA ASC 2023*, pages 846–853. IEEE.

- Phan Minh Dung. 1995. [On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games](#). *Artificial Intelligence*, 77(2):321–357.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630(8017):625–630.
- Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. [Argumentative large language models for explainable and contestable claim verification](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(14):14930–14939.
- Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2025. [M4FC: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset](#). *arXiv preprint arXiv:2510.23508*.
- Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. [AmbiFC: Fact-checking ambiguous claims with evidence](#). *Transactions of the Association for Computational Linguistics*, 12:1–18.
- Ruiling Guo, Xinwei Yang, Chen Huang, Tong Zhang, and Yong Hu. 2025. [CANDY: Benchmarking LLMs’ limitations and assistive potential in Chinese misinformation fact-checking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5724–5758.
- Haorui He, Yupeng Li, Dacheng Wen, Yang Chen, Reynold Cheng, Donglong Chen, and Francis C. M. Lau. 2026. [Debating Truth: Debate-driven claim verification with multiple large language model agents](#). In *Proceedings of the ACM Web Conference 2026*, pages 8851–8861.
- Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. [Decomposition Dilemmas: Does claim decomposition boost or burden fact-checking performance?](#) In *Proceedings of NAACL 2025*, pages 6313–6336.
- Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. [CHEF: A pilot Chinese dataset for evidence-based fact-checking](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376.
- Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. [OpenFactCheck: A unified framework for factuality evaluation of LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. [HoVer: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.
- Avinash Kori, Antonio Rago, and Francesca Toni. 2025. [Free argumentative exchanges for explaining image classifiers](#). In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, pages 1172–1180.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of EMNLP 2024*, pages 17889–17904.
- Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See Kiong Ng, and Tat-Seng Chua. 2025. [FACT-AUDIT: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models](#). In *Proceedings of ACL 2025*, pages 360–381.
- Ying-Jia Lin, Chun-Yi Lin, Chia-Jen Yeh, Yi-Ting Li, Yun-Yu Hu, Chih-Hao Hsu, Mei-Feng Lee, and Hung-Yu Kao. 2024. [CFEVER: A Chinese fact extraction and verification dataset](#). In *Proceedings of AAAI*, volume 38, pages 18626–18634.
- Yuxuan Liu, Hongda Sun, Wenya Guo, Xinyan Xiao, Cunli Mao, Zhengtao Yu, and Rui Yan. 2025. [BiDeV: Bilateral defusing verification for complex claim fact-checking](#). In *Proceedings of AAAI*, volume 39, pages 541–549.
- Xiangwei Lv, Mengze Li, Jingyuan Chen, Zhiang Dong, Sirui Han, and Beishui Liao. 2025. [Out-of-distribution detection via LLM-guided outlier generation for text-attributed graph](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19544–19555.
- Xiucheng Lyu, Chengyu Cao, Mingwei Sun, Bin Liang, Liang Yao, and Ruifeng Xu. 2025. [Bootstrapping LLM-based fact-checking via iterative rationalization finetuning](#). In *ICASSP 2025*, pages 1–5. IEEE.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594.
- Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. [Hallucination-free? assessing the reliability of leading AI legal research tools](#). *Journal of Empirical Legal Studies*, 22(2):216–242.

- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of EMNLP 2023*, pages 12076–12100.
- Raymond S Nickerson. 1998. [Confirmation bias: A ubiquitous phenomenon in many guises](#). *Review of general psychology*, 2(2):175–220.
- Kenji Oki, Kazuhiro Yamashita, and Shinya Kitajima. 2025. [A hybrid approach combining LLMs and web-based information for automated fact-checking](#). In *2025 IEEE International Conference on Web Services (ICWS)*, pages 960–962. IEEE.
- Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. [Fin-Fact: A benchmark dataset for multimodal financial fact-checking and explanation generation](#). In *Companion Proceedings of the ACM on Web Conference 2025*, pages 785–788.
- Lorraine Saju, Arnim Bleier, Jana Lasser, and Claudia Wagner. 2025. [Facts are harder than opinions—a multilingual, comparative analysis of LLM-based fact-checking reliability](#). *arXiv preprint arXiv:2506.03655*.
- 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*, volume 36, pages 65128–65167.
- Vinay Setty. 2024. [Surprising efficacy of fine-tuned transformers for fact-checking over larger language models](#). In *Proceedings of SIGIR 2024*, pages 2842–2846.
- Gautam Kishore Shahi, Amit Kumar Jaiswal, and Thomas Mandl. 2024. [FakeClaim: a multiple platform-driven dataset for identification of fake news on 2023 Israel-Hamas war](#). In *Proceedings of ECIR 2024*, pages 66–74. Springer.
- Neşet Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael J Witbrock. 2023. [Multi2Claim: Generating scientific claims from multi-choice questions for scientific fact-checking](#). In *Proceedings of EACL 2023*, pages 2652–2664.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. [MiniCheck: Efficient fact-checking of LLMs on grounding documents](#). In *Proceedings of EMNLP 2024*, pages 8818–8847.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. [Large language models in medicine](#). *Nature Medicine*, 29(8):1930–1940.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and verification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.
- Andreas Vlachos and Sebastian Riedel. 2014. [Fact checking: Task definition and dataset construction](#). In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.
- William Yang Wang. 2017. [“liar, liar pants on fire”: A new benchmark dataset for fake news detection](#). In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*, pages 422–426.
- Xiaoou Wang, Elena Cabrio, and Serena Villata. 2025. [When automated fact-checking meets argumentation: Unveiling fake news through argumentative evidence](#). *Argument & Computation*, 16(3):405–424.
- Bin Wei, Yaoyao Yu, Leilei Gan, and Fei Wu. 2025. [An LLMs-based neuro-symbolic legal judgment prediction framework for civil cases](#). *Artificial Intelligence and Law*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Zhenyu Cui, Rong Liu, Zhaozhao Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. [FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 137010–137045.
- Xuan Zhang and Wei Gao. 2024. [Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM](#). In *Proceedings of LREC-COLING 2024*, pages 13861–13873.
- Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. [ArgRAG: Explainable retrieval augmented generation using quantitative bipolar argumentation](#). In *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, volume 284 of *Proceedings of Machine Learning Research*, pages 697–718.

## A Abstract Argumentation

This appendix provides the full formal treatment of abstract argumentation, expanding on the core definitions given in Section 4. The main text introduces the AAF tuple  $\mathcal{F} = (\mathcal{A}, \mathcal{R})$  together with the notions of attack, conflict-freeness, defense, and the grounded extension. Here we additionally formalize admissibility and the complete and preferred extensions, which situate grounded semantics within the broader family of argumentation semantics. Recall that for two arguments  $a, b \in \mathcal{A}$ ,  $(a, b) \in \mathcal{R}$  means that  $a$  attacks  $b$ , i.e., the acceptance of  $a$  counts as a reason against accepting  $b$ . Given a candidate set  $B \subseteq \mathcal{A}$  and an argument  $a \in \mathcal{A}$ :

- $B$  attacks  $a$  if there exists  $b \in B$  such that  $(b, a) \in \mathcal{R}$ ;
- $a$  attacks  $B$  if there exists  $b \in B$  such that  $(a, b) \in \mathcal{R}$ .

An argument  $a$  is *defended* by a set  $B \subseteq \mathcal{A}$  iff for every  $b \in \mathcal{A}$  with  $(b, a) \in \mathcal{R}$  there exists  $c \in B$  such that  $(c, b) \in \mathcal{R}$ . For two sets  $B, P \subseteq \mathcal{A}$ , we say that  $B$  attacks  $P$  if there exist  $a \in B$  and  $b \in P$  such that  $(a, b) \in \mathcal{R}$ .

**Definition 2** (Argumentation Semantics). *Let  $\mathcal{F} = (\mathcal{A}, \mathcal{R})$  be an abstract argumentation framework and  $B \subseteq \mathcal{A}$ .*

- $B$  is conflict-free iff there are no  $a, b \in B$  such that  $(a, b) \in \mathcal{R}$ .
- $B$  is admissible iff  $B$  is conflict-free and every  $a \in B$  is defended by  $B$ .
- $B$  is a complete extension iff  $B$  is admissible and, for every  $a \in \mathcal{A}$ ,  $B$  defends  $a$  implies  $a \in B$ .
- $B$  is the grounded extension iff  $B$  is the  $\subseteq$ -minimal complete extension.
- $B$  is a preferred extension iff  $B$  is a  $\subseteq$ -maximal admissible set.

These semantics form a hierarchy: the grounded extension is complete, and every complete extension is admissible. Grounded and preferred semantics sit at two ends of this spectrum, corresponding respectively to a skeptical (cautious) and a credulous way of accepting arguments. The grounded extension is unique and always exists, whereas a framework may admit several preferred extensions. ArgCheck adopts grounded semantics throughout, for the reasons given in Section 4. The other semantics are included here only to position this choice within the standard theory.

## B Dataset Details

### B.1 Dataset Comparison

Table 3 compares MAVEN with representative fact-checking datasets across key dimensions, including language coverage, dataset size, annotation schema, and claim authenticity. Existing benchmarks often provide verdict labels and evidence, but many lack ruling-style explanations or systematic error categorization. In contrast, MAVEN introduces an eight-field annotation schema covering verdicts, evidence sources with credibility ratings, ruling-style explanations, corrected claim, topic labels, and a two-level error taxonomy. These annotations support evaluation beyond verdict accuracy, including explanation quality and error pattern analysis.

Synthetic benchmarks such as FEVER (Thorne et al., 2018) and HoVer (Jiang et al., 2020) generate claims from structured knowledge bases. In contrast, MAVEN draws on claims from social media platforms and professional fact-checking websites. This design better reflects real-world misinformation patterns and multi-source evidence conflicts. MAVEN contains both Chinese and English claims across 10 domains, supporting cross-domain and cross-lingual evaluation.

### B.2 Dataset Statistics

Figure 8 illustrates the distribution of MAVEN across three dimensions. Figure 8(a) shows the verdict distribution: 43.8% Refuted, 42.2% Supported, and 14.0% Not Enough Evidence (NEE). This distribution provides a near-balanced split between supported and refuted claims, together with a smaller but substantial NEE subset for diagnostic evaluation. Figure 8(b) presents the distribution of first-level error types among Refuted and NEE claims. Unverifiable Claims form the largest category (166 instances), followed by Evidence Errors (140), Semantic Errors (120), Factual Errors (99), Temporal Errors (88), and Data Errors (72). This distribution shows that unverifiability, evidence quality, and semantic reasoning are major challenges in real-world fact-checking, motivating the need for structured evidence analysis and conflict resolution. Figure 8(c) displays the topic distribution across 10 domains. Entertainment (184 claims), Economy (156), Society (152), Sports (147), and Politics (146) are the largest categories, while Science (112), Health (87), Policy (77), Culture (76), and Environment (48) provide additional domain coverage. This distribution sup-

| Dataset                               | Language     | Size         | Verdict        | Topics    | Evidence | Expl.    | Err.                 | Real     |
|---------------------------------------|--------------|--------------|----------------|-----------|----------|----------|----------------------|----------|
| FEVER (Thorne et al., 2018)           | EN           | ~185K        | 3-class        | Open      | ✓        | ✗        | ✗                    | ✗        |
| AmbiFC (Glockner et al., 2024)        | EN           | ~10.7K       | 3-class        | Open      | ✓        | ✓        | ✗                    | ✓        |
| Fin-Fact (Rangapur et al., 2025)      | EN           | 3,369        | 3-class        | 1         | ✓        | ✓        | ✗                    | ✓        |
| LIAR (Wang, 2017)                     | EN           | ~12.8K       | 6-class        | 1         | ✗        | ✗        | ✗                    | ✓        |
| AVeriTeC (Schlichtkrull et al., 2023) | EN           | ~4.6K        | 4-class        | Open      | ✓        | ✓        | ✗                    | ✓        |
| Med-Fact (Tan et al., 2023)           | EN           | 150K         | 3-class        | 1         | ✓        | ✗        | ✗                    | ✗        |
| M4FC (Geng et al., 2025)              | Multi        | 6,980        | 2-class        | Open      | ✓        | ✗        | ✗                    | ✓        |
| FakeClaim (Shahi et al., 2024)        | Multi        | ~1.5K        | 1-class        | 1         | ✓        | ✗        | ✗                    | ✓        |
| FactSpan (Saju et al., 2025)          | Multi        | ~61.5K       | 5-class        | 5         | ✗        | ✗        | ✗                    | ✓        |
| CFEVER (Lin et al., 2024)             | ZH           | ~30K         | 3-class        | 11        | ✓        | ✗        | ✗                    | ✗        |
| CHEF (Hu et al., 2022)                | ZH           | 10K          | 3-class        | 5         | ✓        | ✗        | ✗                    | ✓        |
| CANDY (Guo et al., 2025)              | ZH           | ~20K         | 2-class        | 7         | ✓        | ✓        | ✗                    | ✓        |
| <b>MAVEN</b>                          | <b>ZH+EN</b> | <b>1,185</b> | <b>3-class</b> | <b>10</b> | <b>✓</b> | <b>✓</b> | <b>✓<sup>†</sup></b> | <b>✓</b> |

Table 3: Comparison of representative fact-checking datasets. Expl. and Err. denote explanation and claim-level error-taxonomy annotations, respectively, and Real denotes real-world claims. Topics reports the number of broad topic domains, while Open indicates open-domain coverage without a fixed domain count. ✓ indicates that a feature is explicitly annotated, whereas ✗ indicates that it is absent. Real-world claims originate from social media, news outlets, or fact-checking platforms. <sup>†</sup>MAVEN provides a two-level claim-error taxonomy consisting of error types and subtypes.

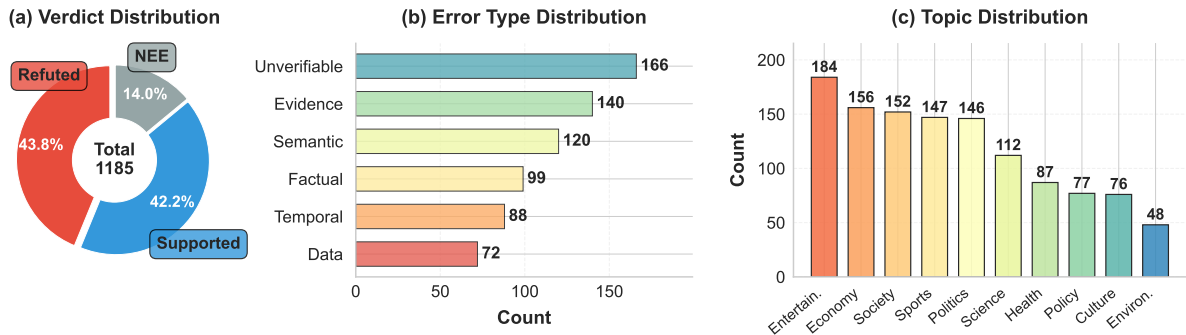


Figure 8: Distribution analysis of the MAVEN dataset across three dimensions: (a) verdict labels, (b) error types, and (c) topic categories.

ports evaluation across diverse reasoning requirements and knowledge domains.

### B.3 Error Taxonomy

MAVEN employs a two-level error taxonomy to support fine-grained analysis of misinformation patterns. Table 4 presents the complete taxonomy with definitions. The first level comprises six error types: (1) **Factual Errors** involve incorrect objective facts that contradict verifiable information, including entity attribute errors, public fact errors, specialized knowledge errors, misattribution, and event misidentification; (2) **Temporal Errors** cover incorrect time information, outdated data, or dynamic event misrepresentation; (3) **Data Errors** include statistical inaccuracies and erroneous quantitative comparisons; (4) **Semantic Errors** capture reasoning failures such as causal confusion, conditional

misuse, scope confusion, and concept confusion; (5) **Evidence Errors** identify fabricated sources, unverified reports, missing evidence chains, or out-of-context interpretations; and (6) **Unverifiable Claims** cover claims lacking accessible authoritative evidence or involving controversies, predictions, or unresolved disputes. The second level provides finer subcategories for each error type, enabling both coarse-grained error analysis and more detailed subcategory-level diagnosis. This hierarchical structure helps identify recurring failure patterns and supports targeted improvements in fact-checking methods.

### B.4 Annotation Details

The dataset was annotated by two graduate-level annotators with backgrounds in NLP and bilingual proficiency in Chinese and English. All annota-

| Error Type          | Subcategory                | Definition                                                                 |
|---------------------|----------------------------|----------------------------------------------------------------------------|
| <b>Factual</b>      | Entity Attribute           | Incorrect attributes of persons, organizations, or locations.              |
|                     | Public Fact                | Contradicting publicly verifiable facts.                                   |
|                     | Specialized Knowledge      | Misusing domain-specific knowledge (medicine, law, physics).               |
|                     | Misattribution             | Incorrectly attributing statements, actions, or achievements.              |
|                     | Event Misidentification    | Misidentifying an entire event as another event.                           |
| <b>Temporal</b>     | Incorrect Time             | Wrong time points or temporal sequences.                                   |
|                     | Outdated Information       | Using superseded data or policies.                                         |
|                     | Dynamic Event              | Treating evolving events as static facts.                                  |
| <b>Data</b>         | Statistical Data           | Numerical inaccuracies in statistics.                                      |
|                     | Data Comparison            | Erroneous quantitative comparisons.                                        |
| <b>Semantic</b>     | Causal Reasoning           | Confusing correlation with causation or reversing causal direction.        |
|                     | Conditional                | Misusing conditional logic (sufficient vs. necessary conditions).          |
|                     | Scope Confusion            | Overgeneralizing from partial to whole.                                    |
|                     | Concept Confusion          | Conflating distinct but related concepts.                                  |
| <b>Evidence</b>     | Fake Sources               | Fabricated or non-existent sources.                                        |
|                     | Unverified Reports         | Asserting an unfounded rumor, leak, or unconfirmed report as fact.         |
|                     | Missing Evidence Chain     | Lack of supporting evidence for the claim.                                 |
|                     | Out-of-context             | Misinterpreting evidence by removing context.                              |
| <b>Unverifiable</b> | Lack of Authoritative Data | No accessible authoritative record exists to confirm or refute the claim.  |
|                     | Controversy                | Claims involving ongoing debates without consensus.                        |
|                     | Future Predictions         | Future events, forecasts, subjective predictions, or unconfirmed plans.    |
|                     | Conflicting Sources        | Authoritative sources provide contradictory statements without resolution. |

Table 4: Two-level error taxonomy in MAVEN with definitions.

tors were trained on the annotation manual prior to labeling and followed a shared schema and verdict-definition protocol. Each instance was assigned a verdict, a justification, up to three credibility-rated evidence sources, a corrected claim, and a topic; Refuted and NEE instances additionally received a two-level error annotation, following the procedure described in Section 3. Discrepancies between the generated and pre-labeled verdicts triggered manual review, while consistent cases underwent spot-checking. The complete annotation manual, including the schema, credibility criteria, and boundary-case rules, is provided in our repository.<sup>1</sup>

## C Detailed Performance and Robustness Analysis

This section supplements the main results with a cross-backbone comparison, a human audit of conflict detection, and a pilot computational cost comparison among the evaluated inference settings.

### C.1 Cross-Backbone Evaluation

The main experiments use qwen3-max as the backbone model, raising the question of whether the observed improvements are specific to this model. To examine the cross-backbone robustness of ArgCheck, we replace qwen3-max with deepseek-v4-pro and rerun the five baselines and ArgCheck on a stratified subset of 200 instances

| Method          | Accuracy    | Macro-P     | Macro-R     | Macro-F1    |
|-----------------|-------------|-------------|-------------|-------------|
| ZS              | 51.0        | 55.4        | 48.3        | 46.7        |
| ZS-T            | 53.0        | 56.8        | 49.8        | 48.4        |
| CoT             | 53.5        | 56.2        | 51.1        | 48.2        |
| SR              | 62.5        | 58.4        | 53.7        | 53.3        |
| MAD             | 60.0        | 57.6        | 54.8        | 53.7        |
| <b>ArgCheck</b> | <b>75.5</b> | <b>64.2</b> | <b>61.2</b> | <b>62.1</b> |

Table 5: Verdict-prediction performance with DeepSeek-V4-Pro on a label-stratified 200-claim subset of AVeriTeC (dev). Results are reported as percentages, with the best results **highlighted**.

from the AVeriTeC development set. The subset preserves the label distribution of the complete development set and is sampled with seed 42. For each method, we retain the prompting, retrieval, and inference configurations used in the main experiments, and set the maximum number of rounds to  $R = 5$  for the iterative methods.

As shown in Table 5, ArgCheck achieves the best performance across all four metrics, reaching 75.5% accuracy, 64.2 Macro-Precision, 61.2 Macro-Recall, and 62.1 Macro-F1. Compared with the strongest baseline for each metric, ArgCheck improves accuracy by 13.0 percentage points over Self-Refine and Macro-F1 by 8.4 points over Multi-Agent Debate. Its improvements in both Macro-Precision and Macro-Recall indicate that the over-

| Predicted Label | Pairs | Agreement Rate |
|-----------------|-------|----------------|
| Conflict        | 100   | 75.0%          |
| Non-conflict    | 100   | 86.0%          |

Table 6: Human audit of the LLM-based conflict detector on 200 evidence pairs, comprising 100 pairs predicted as conflicting and 100 pairs predicted as non-conflicting. The class-specific agreement rates are conditional on the predicted labels.

all gain is not attributable solely to one side of the precision–recall trade-off. Although this experiment is conducted on a 200-instance subset rather than the complete development set, the consistent improvement provides preliminary evidence that the effectiveness of the argumentative reasoning framework is not specific to Qwen3-Max and can transfer to a different LLM backbone.

### C.2 Human Audit of Conflict Detection

ArgCheck uses the LLM-based predicate  $cf(\cdot, \cdot)$  to determine whether two pieces of evidence are in conflict. Because errors in conflict detection may alter the attack relations in the argumentation framework and subsequently affect the grounded extension, we conduct a human audit of the detector. We sample 200 evidence pairs stratified according to the predictions of the detector, including 100 pairs predicted as conflicting and 100 predicted as non-conflicting, using a fixed random seed of 42. A human annotator, blind to the model predictions, independently determines whether each pair contains a substantive conflict.

As shown in Table 6, the human annotator agrees with 75 of the 100 predicted conflicts and 86 of the 100 predicted non-conflicts, yielding an overall agreement of 80.5% on the balanced audit set. Among the disagreements, 25 pairs predicted as conflicting are judged non-conflicting by the annotator. These cases often involve evidence that discusses the same claim from different perspectives without making logically incompatible assertions. Conversely, 14 pairs predicted as non-conflicting are judged to contain a conflict, often because the contradiction is expressed implicitly and requires contextual or pragmatic inference. These results show that the detector correctly captures a substantial majority of the audited relations while still exhibiting non-negligible over-detection and under-detection. Since the sample is stratified by the model prediction and evaluated by a single annotator, the results should be interpreted as an analysis

| Setting              | API Calls | Tokens | Latency (s) |
|----------------------|-----------|--------|-------------|
| Zero-Shot            | 1.0       | 1.1K   | 2.0         |
| Self-Refine          | 1.0       | 1.8K   | 4.8         |
| Multi-Agent Debate   | 14.0      | 70.2K  | 59.6        |
| ArgCheck ( $R = 2$ ) | 20.7      | 65.5K  | 44.0        |
| ArgCheck ( $R = 5$ ) | 47.0      | 188.7K | 100.7       |

Table 7: Pilot computational cost analysis on 10 stratified AVeriTeC development instances. All values are averaged per claim. Tokens denote the sum of LLM input and output tokens, and latency is measured under parallel execution.

of representative error patterns rather than an estimate of corpus-level detection performance.

### C.3 Pilot Computational Cost Analysis

ArgCheck introduces additional model and retrieval calls through iterative adversarial evidence collection and argumentation-framework construction, particularly for targeted counter-evidence retrieval and LLM-based conflict detection. To provide a pilot estimate of the associated computational cost, we evaluate five method configurations on 10 instances from the AVeriTeC development set. The sample is allocated proportionally across verdict labels using the largest-remainder rule and drawn with seed 13. For each setting, we record the total number of API calls, the sum of LLM input and output tokens, and the wall-clock latency. The API count includes both LLM calls and standalone retrieval calls, whereas search integrated into an LLM API call is counted as part of that call rather than as a separate retrieval request. Latency is measured as the elapsed wall-clock time per instance under the parallel execution strategy used in our implementation. All quantities are averaged over the 10 sampled instances.

Table 7 shows that MAD and ArgCheck incur higher inference costs than the single-agent baselines. Reducing the maximum number of ArgCheck rounds from  $R = 5$  to  $R = 2$  decreases the average number of API calls from 47.0 to 20.7, token usage from 188.7K to 65.5K, and latency from 100.7 to 44.0 seconds per claim, corresponding to reductions of approximately 56%, 65%, and 56%, respectively. Compared with MAD, ArgCheck with  $R = 2$  makes more API calls but uses approximately 6.7% fewer tokens and achieves 26.2% lower wall-clock latency under parallel execution. Together with the round ablation reported in the main text, these results indicate that  $R = 2$  offers a lower-cost ArgCheck configuration

while retaining most of the performance gain.

## D Authority and Timeliness Scoring

To construct attack relations in the argumentation framework, we assign each argument  $a$  two scores: an authority score  $\alpha(a)$  and a timeliness score  $\tau(a)$ . These scores are used only when two arguments are in semantic conflict.

### D.1 Authority Scoring

The authority score  $\alpha(a)$  reflects the credibility of the evidence source. We classify sources into three levels:

- **High Authority** ( $\alpha = 3$ ): Government agencies, international organizations (e.g., WHO, IMF), authoritative databases, and academic publications.
- **Medium Authority** ( $\alpha = 2$ ): Industry reports, professional blogs, regional mainstream news outlets, and established media organizations.
- **Low Authority** ( $\alpha = 1$ ): Self-media platforms, anonymous accounts, and unverified social media content.

The authority level is automatically determined using a curated domain whitelist for each category. For example, domains such as .gov and .edu are classified as high authority, while personal blogs and social media platforms receive lower ratings.

### D.2 Timeliness Scoring

The timeliness score  $\tau(a)$  captures the recency of the evidence. This is useful for dynamic or time-sensitive claims, where later evidence may supersede earlier reports. Given the publication timestamp  $t(a)$  of the evidence source and a fixed reference time  $t_{\text{ref}}$  (the retrieval date used for the experimental run), we compute the time difference in days:

$$\Delta t(a) = t_{\text{ref}} - t(a). \quad (6)$$

The timeliness score is calculated using an exponential decay function:

$$\tau(a) = e^{-\lambda \cdot \Delta t(a)}. \quad (7)$$

Here,  $\lambda$  controls how quickly evidence loses timeliness. We set  $\lambda = \frac{\ln 2}{365}$ , giving a half-life of one year. Thus, evidence published one year before  $t_{\text{ref}}$  receives half the timeliness score of evidence published at  $t_{\text{ref}}$ . This function gives more recent evidence higher scores while allowing older evidence to decay smoothly without becoming zero. If a publication timestamp is unavailable, we assign the lowest timeliness score so that undated evidence is not favored by recency.

| $\alpha(a), \alpha(b)$ | $\tau(a)$ vs. $\tau(b)$ | Outcome               | Rationale                     |
|------------------------|-------------------------|-----------------------|-------------------------------|
| 3, 2                   | $\tau(a) \geq \tau(b)$  | $a \rightarrow b$     | higher authority; not older   |
| 2, 2                   | $\tau(a) > \tau(b)$     | $a \rightarrow b$     | equal authority; $a$ is newer |
| 3, 2                   | $\tau(a) < \tau(b)$     | $a \leftrightarrow b$ | authority–recency trade-off   |
| 2, 2                   | $\tau(a) = \tau(b)$     | $a \leftrightarrow b$ | tie on both criteria          |

Table 8: Worked examples of the attack-orientation predicate. A directed attack  $a \rightarrow b$  is added whenever  $\pi(a, b)$  holds. Mutual attacks arise when both  $\pi(a, b)$  and  $\pi(b, a)$  hold, including cases where the arguments are higher on different criteria or tied on both.

### D.3 Combined Priority in Attack Relations

For two arguments  $a, b \in \mathcal{A}$  that share a sub-claim, an ordered attack from  $a$  to  $b$  is added when the symmetric conflict predicate  $\text{cf}(a, b)$  holds and the attack-orientation predicate  $\pi(a, b)$  also holds:

$$(a, b) \in \mathcal{R} \Leftrightarrow \text{cf}(a, b) \wedge \pi(a, b). \quad (8)$$

The predicate  $\pi(a, b)$  holds iff  $\alpha(a) > \alpha(b)$ ,  $\tau(a) > \tau(b)$ , or  $\alpha(a) = \alpha(b) \wedge \tau(a) = \tau(b)$ . Because  $\pi(a, b)$  is evaluated for each ordered pair, a single inequality does not necessarily yield a one-way edge: if  $a$  is higher in authority but  $b$  is more recent, both  $\pi(a, b)$  and  $\pi(b, a)$  hold, producing mutual attacks. A one-directional attack is added only when one argument is no lower on both criteria and strictly higher on at least one; ties on both criteria also yield mutual attacks. This design avoids manually fixing a trade-off between authority and recency when the two criteria disagree. In such cases, acceptance is determined by grounded semantics and the surrounding defense structure rather than by a hand-coded ranking rule. Table 8 illustrates each case with a worked example.

## E System Prompts

This appendix presents the key prompts used in ArgCheck. The prompts are formatted for readability, with substantive instructions preserved.

### E.1 Claim Decomposition Prompt

**Purpose:** Decompose a complex claim into atomic, verifiable sub-claims.

**Prompt:**

You are a fact-checking expert. Your task is to decompose a complex claim into verifiable sub-claims.

**\*\*Original Claim:\*\***  
{claim}

**\*\*Task:\*\***  
Analyze this claim and break it down into atomic sub-claims that need to be verified independently.

**\*\*Decomposition Principles:\*\***

- \*\*Atomic\*\*:** Each sub-claim should be a single, independent assertion
- \*\*Verifiable\*\*:** Each sub-claim can be verified through evidence search
- \*\*Complete\*\*:** Together, all sub-claims should cover the full scope of the original claim
- \*\*Non-Redundant\*\*:** Do not repeat the same fact in multiple sub-claims.

**\*\*Sub-Claim Types:\*\***

- \*\*Entity\*\*:** Who/what is the main subject? (e.g., "X is a person/organization/thing")
- \*\*Event\*\*:** Did something happen? (e.g., "Event X occurred")
- \*\*Time\*\*:** When did it happen? (e.g., "Event happened in 2020")
- \*\*Quantity\*\*:** How much/many? (e.g., "X has 23 Grand Slam titles")
- \*\*Qualifier\*\*:** Scope/degree modifiers (e.g., "most in history" vs "most in Open Era")
- \*\*Relationship\*\*:** Connection between entities (e.g., "X is CEO of Y")

**\*\*Output Format (JSON only, no other text):\*\***  
{ "sub\_claims": [{ "id": "c1", "text": "..."}, { "id": "c2", "text": "..."}] }

**\*\*IMPORTANT\*\*:** Return ONLY valid JSON. Do not include any text before or after the JSON object.

- Focus on the MAIN assertions - don't over-decompose

Now decompose the claim:

## E.2 Pro / Con Agent Query Generation Prompts

### E.2.1 First Round Query Generation

**Purpose:** Generate a concise search query for a target sub-claim.

**Prompt:**

You are a fact-checking expert for the {stance} side, responsible for finding {goal} evidence for the following claim.  
Claim: {claim}  
Target sub-claim: {sub\_claim}

Task objective:  
Generate exactly ONE stance-conditioned search query to find {goal} evidence for the target sub-claim.

Analysis steps:

1. Focus on the target sub-claim, not the whole parent claim
2. Preserve critical entities, numbers, timeframes, and qualifiers
3. Prefer official, authoritative, or primary-source wording

Query generation requirements:

1. Precision: Target the ONE most important verification point of the sub-claim
2. Conciseness: Keep the query SHORT and focused on keywords (max 50 characters for Chinese, 100 for English)
3. Authority: Prioritize locating credible sources such as official websites, government agencies, academic journals, authoritative media, etc.
4. Format: Use keywords or short phrases, NOT full questions

Query format specifications:

- Use keywords connected with plus signs (+), following the same length limit above
- Good: 'Pluto+aphelion+distance+AU'

Output requirements:

Output exactly ONE SHORT search query (keywords/short phrase). Single line. Do not include explanations, question marks, numbering, or extra text.

The language used for the query should be consistent with the claim.

Now please provide the query:

### E.2.2 Follow-up Round Query Generation

**Purpose:** Counter opponent's evidence in subsequent rounds.

**Prompt:**

You are a fact-checking expert for the {stance} side, responsible for finding {goal} evidence for the following claim.  
Claim: {claim}  
All sub-claims (pick the ONE most worth rebutting on this round):  
{sub\_claims\_block}  
Current round: Round {round\_num}  
Opponent evidence found in the previous round:  
{opponent\_evidence}  
Already-searched queries (avoid duplicating):  
{existing\_queries}

Task objective:

1. Pick exactly ONE sub-claim where the opponent's evidence is strongest or most threatening to your side.
2. Generate ONE precise search query that targets the weak point of the opponent's evidence on that sub-claim, in order to find authoritative evidence for the {stance} side.

Query generation requirements:

1. Precision: The query should directly correspond to the core content of the claim
2. Conciseness: Keep query SHORT and focused on keywords

3. Authority: Prioritize locating credible sources
4. Targeting: If opponent evidence exists, target its weak points
5. Differentiation: Avoid repetition with already searched topics

Query format specifications:  
- Use keywords/short phrases

Output format (CRITICAL - exactly two lines, nothing else):  
SUB\_CLAIM\_ID: <id of the chosen sub-claim>  
QUERY: <the keyword query>

The language used for the query should be consistent with the claim.

### E.3 Content Conflict Detection Prompt

**Purpose.** Detect semantic conflicts between same-sub-claim argument pairs before priority-based attack edges are added.

#### Prompt:

You are a fact-checking expert specializing in argument analysis.

Task:

Identify semantic conflicts only among the candidate pairs below. A conflict exists if two arguments make contradictory claims, present contradictory evidence, rely on mutually exclusive premises, or one directly refutes the other for the same sub-claim.

Only decide whether a symmetric semantic conflict exists; do not determine attack direction or priority.

**\*\*Original Claim:\*\***  
{claim}

Candidate pairs:  
{candidate\_pairs}

Output one line for each conflict:  
CONFLICT: <A1\_id> vs <A2\_id> | <conflict\_type>

Use conflict\_type from:  
contradictory\_conclusion,  
contradictory\_evidence, logical\_opposition,  
direct\_refutation.

If there are no conflicts, output exactly:  
NO\_CONFLICTS

Now analyze:

You are an impartial fact-checking judge. Your task is to adjudicate the following claim using the accepted arguments as primary evidence, while using all arguments and attack relations to account for defeated or disputed evidence.

Claim: "{claim}"

**\*\*Accepted Arguments:\*\***  
{accepted\_arguments}

**\*\*Verdict Definition:\*\***

1. **\*\*Supported\*\***: there is sufficient evidence to show that the claim is supported by the evidence.
2. **\*\*Refuted\*\***: there is sufficient evidence to refute the claim. OR the claim asserts something that should appear in authoritative sources (e.g., public figure's statement, official policy, government data, major public event), and yet no relevant record can be found in such sources.
3. **\*\*Not Enough Evidence\*\***: there is insufficient / conflicting / ambiguous evidence to either support or refute the claim.

TIPS:

1. If the evidence is entirely about a different topic, this is NOT refutation, it is Not Enough Evidence.
2. If no relevant evidence was retrieved, AND the claim is not the type expected to appear in authoritative sources (e.g., private events, folk rumors, predictions), it is Not Enough Evidence.

**\*\*All Arguments:\*\***  
{all\_arguments}

**\*\*Attack Relations:\*\***  
{attack\_relation}

Output format (JSON):

```
{
  "justification": "Reasoning in {response_language}. Explain how accepted arguments support, refute, or fail to decide the original claim, and use all arguments and attack relations to explain accepted, defeated, or unresolved evidence.",
  "verdict": "Supported" / "Refuted" / "Not Enough Evidence",
  "confidence": "<a number between 0 and 1>",
  "key_argument_ids": ["key argument id list"]
}
```

Return only valid JSON. Now make the verdict:

### E.4 Judge Verdict Generation Prompt

#### E.4.1 Benchmark: MAVEN

**Purpose.** Generate the final verdict from accepted arguments and the full argumentation framework for MAVEN.

#### Prompt:

#### E.4.2 Benchmark: AVeriTeC

**Purpose.** Generate the final verdict from accepted arguments and the full argumentation framework for AVeriTeC.

## Prompt:

You are an impartial fact-checking judge. Your task is to adjudicate the claim using the accepted arguments as primary evidence, while using all arguments and attack relations to account for defeated or disputed evidence.

Claim: "{claim}"

Accepted Arguments:  
{accepted\_arguments}

**\*\*Verdict Definition\*\*:**

1. Supported:

The claim is supported by the accepted arguments. Use this label when the available evidence is sufficient to establish the claim's main factual content.

2. Refuted:

The claim is contradicted by the accepted arguments. Use this label when the available evidence directly contradicts the claim's main factual content.

3. Not Enough Evidence:

- There are not enough accepted arguments to support or refute the claim. Use this label when appropriate evidence cannot be found, or when some important aspect of the claim remains neither supported nor refuted.
- The claim is misleading due to unresolved conflicts or cherry-picking, but is not explicitly refuted. This includes true-but-misleading claims, selective presentation of evidence, omitted context, or cases where credible evidence in the full framework materially qualifies the accepted evidence.

Boundary rules:

- If one component of a multi-part claim is refuted and another component has not enough evidence, label the whole claim Refuted.
- If one component is supported but another important component has not enough evidence, label the whole claim Not Enough Evidence.

Graph-use rule:

- Base the verdict primarily on accepted arguments. Use all arguments and attack relations to explain defeated or unresolved conflicts, but do not treat defeated arguments as final evidence.

All arguments in the argumentation framework:  
{all\_arguments}

Attack relations:  
{attack\_relation}

Output format (JSON):

```
{{
  "justification": "Reasoning in
    {response_language}. Explain how accepted
    arguments support, refute, or fail to
    decide the original claim, and use all
    arguments and attack relations to explain
```

```
    accepted, defeated, or unresolved
    evidence.",
  "verdict": "<one of: {verdict_options}>",
  "confidence": "<a number between 0 and 1>",
  "key_argument_ids": ["key argument id list"]
}}
```

Return only valid JSON. Now make the verdict: