STRUCTURED EVIDENCE FILTERING AND CONSENSUS REASONING FOR MULTI-AGENT FACT-CHECKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale retrieval-augmented fact-checking often suffers from noisy evidence and unreliable multi-agent reasoning aggregation. To address these challenges, we propose a Structured Evidence Filtering-based Multi-Agent Consensus Framework (**SEF-MAC**), consisting of two core modules. The first is a *Structured Evidence Filtering Module* (SEF), which leverages domain knowledge to remove irrelevant evidence and maintain semantic consistency with the target claim. The second is a *Challenge-and-Validation Reasoning Approach*, which treats minority dissent as a signal to activate iterative cross-examination, allowing agents to correct logical defects through evidence-grounded deliberation. Experiments on three benchmarks—CheckThat-T3, CHEF, and SciFact—show that SEF-MAC consistently outperforms competitive baselines, with Micro-F1 gains of up to 3.72% and Macro-F1 gains of up to 3.00%. Further ablation and analysis suggest that evidence filtering and consensus reasoning make complementary contributions to improve fact-checking accuracy across different evidence settings.

1 INTRODUCTION

In the era of large language models, retrieval-augmented automated fact-checking has become pivotal for verifying real-world factual claims and misinformation (Wadden et al., 2020). Existing approaches typically retrieve external evidence from search results or domain-specific corpora, and verify claims through evidence-based models (Hu et al., 2023), planning-based frameworks (Zhao et al., 2024), agent-based systems (Shukla et al., 2025), or multi-agent debate frameworks (Han et al., 2025; Lakara et al., 2024).

Yet mainstream methods still face two key challenges: *noisy evidence* and *unreliable reasoning* (Hazarika et al., 2025), especially in complex and low-evidence scenarios that require both high-quality evidence and rigorous logical inference (Kolli et al., 2025; Voinea, 2025). First, many systems rely heavily on retrieved search results without sufficiently filtering irrelevant or misleading evidence (Xue et al., 2025; Upravitelev et al., 2025). Second, existing reasoning frameworks may lack robust logical consistency and effective self-correction; in multi-agent settings, this can make it difficult to form well-supported collective conclusions (Han et al., 2025) and may lead to hallucinated or biased reasoning under incomplete evidence (Kolli et al., 2025; Voinea, 2025).

To address these gaps, we propose SEF-MAC, a unified multi-agent framework that combines *evidence filtering* and *reasoning consensus* for retrieval-augmented fact-checking. Specifically, we introduce a *Structured Evidence Filtering Module* that assesses retrieved snippets through structured factual consistency checks against the target claim, removing noisy and irrelevant evidence while retaining information that is relevant and claim-consistent. This process provides more focused evidence for subsequent verification, reducing the interference of unreliable retrieval results.

To further enhance reasoning reliability and accuracy, we design a *Challenge-and-Validation Reasoning Approach*. Unlike naive multi-agent ensembling or simple majority voting, it treats minority dissent as a useful signal to activate iterative deliberation. When an agent identifies insufficient evidence, inconsistent reasoning, or a potential logical flaw, the framework initiates a cross-examination process among agents. Through evidence-grounded challenge and validation, this module is designed to surface possible logical gaps, reconcile divergent views, and improve the identification of evidence-deficient claims.

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.

Experiments conducted on CheckThat-T3 (Alam et al., 2025), CHEF (Hu et al., 2022), and SciFact (Wadden et al., 2020) verify that SEF-MAC achieves improved performance over competitive baselines across all three benchmarks. Combining evidence filtering and consensus reasoning provides a promising approach for complex fact-checking tasks.

Our contributions are summarized as follows:

- We propose SEF-MAC, a unified multi-agent fact-checking framework designed to address noisy evidence and unreliable reasoning consensus in retrieval-augmented fact-checking. It tightly couples evidence filtering with collaborative reasoning to support better-grounded verification under complex and evidence-insufficient scenarios.
- We propose two mechanisms for noisy fact-checking: SEF scores each candidate paragraph along factual relevance, temporal consistency, entity alignment, and thematic coherence for adaptive filtering; and Challenge-and-Validation uses non-unanimous verdicts as a deliberation trigger, enabling targeted validation instead of simple majority voting or always-on debate.
- We evaluate the effectiveness of the proposed method on three benchmarks. The experimental results demonstrate SEF-MAC consistently outperforms competitive baselines, achieving Micro-F1 gains of up to 3.72% and Macro-F1 gains of up to 3.00%.

2 Related Work

2.1 Retrieval-Augmented Fact-Checking

Retrieval-Augmented Generation (RAG) has become a widely used paradigm for automated fact-checking (Ge et al., 2025; Russo et al., 2025). It verifies claims by coupling evidence retrieval with logical reasoning (Dharmavaram & Hakak, 2025). Frameworks like Self-RAG and ReACT use reflection (Fadeeva et al., 2025) to improve quality. Although these methods enhance reasoning faithfulness, they still decouple evidence filtering from final verification (Liu et al., 2026; Rolinger & Liu, 2025). This separation makes it difficult to capture sparse valid signals under high noise (Shukla et al., 2025), and single-agent models often miss key evidence buried in irrelevant content (Tang et al., 2025). Unlike these pipelines, SEF-MAC couples evidence filtering with collaborative reasoning by forwarding a compact, claim-consistent evidence set to role-specialized verifier agents, so that downstream consensus is formed over denoised evidence rather than raw retrieval results.

2.2 Evidence Pruning and Selection

Evidence quality directly determines reasoning accuracy (Peng et al., 2026). Studies show that low-quality evidence can severely degrade performance in noisy environments (Bussotti & Papotti, 2025; Zeng et al., 2025). Although existing reranking strategies improve relevance coverage (Song et al., 2025), they seldom leverage domain knowledge to evaluate the logical validity of evidence snippets. For example, FLARE optimizes retrieval timing but fails to filter segments without direct logical connections to the claim (Arakelyan et al., 2025). To overcome these weaknesses, our Structured Evidence Filtering Module goes beyond relevance-oriented reranking by scoring each candidate snippet with structured factual consistency checks, including core fact relevance, temporal consistency, entity alignment, and thematic coherence, before constructing the evidence set for final verification.

2.3 Multi-Agent Collaboration and Consensus

Multi-agent systems such as PACAR (Zhao et al., 2024) and DebateCV (He et al., 2026) improve complex reasoning via specialized coordination. More recent works extend this line along two complementary directions: D2D (Han et al., 2025) reformulates misinformation detection as a structured multi-agent debate to enhance decision transparency, while FIRE (Xie et al., 2025) performs fact-checking via an agent-based iterative loop that alternates between evidence retrieval and verification for more cost-effective reasoning. Although these frameworks have reported performance gains, they can still struggle with complex and evidence-insufficient claims (Xiao et al., 2026). Additionally, conventional voting-based consensus may discard informative minority judgments when the majority prediction is incorrect (Ma et al., 2025; Lee & Lee, 2026). In contrast, the proposed

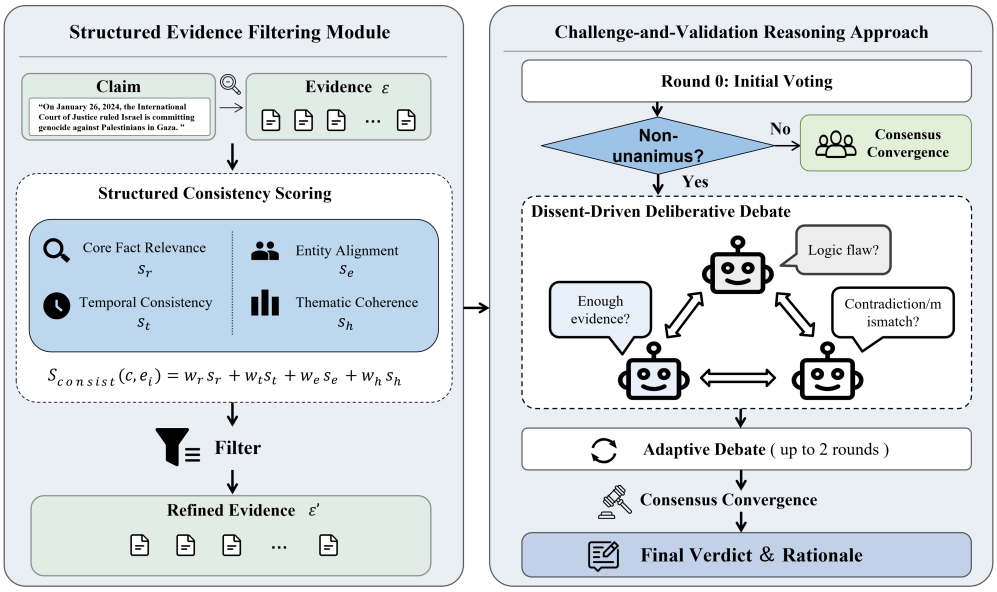

Figure 1: Overview of SEF-MAC. Left: the Structured Evidence Filtering (SEF) module computes four-dimensional consistency scores to filter noisy evidence and produce a refined evidence set. Right: the Challenge-and-Validation Reasoning Approach runs initial voting among three role-specialized verifier agents; non-unanimous verdicts trigger at most two rounds of dissent-driven deliberative debate to derive the final fact-checking verdict and rationale.

Challenge-and-Validation Reasoning Approach takes minority dissent as a key signal to activate targeted cross-examination among agents focused on evidence sufficiency, contradiction detection, and reasoning validity. Such evidence-driven deliberation can help identify possible logical gaps and provide additional validation when the initial agent judgments disagree.

3 Our Framework: SEF-MAC

3.1 Task Definition

Given a claim c and candidate evidence $\mathcal{E} = \{e_1, e_2, \ldots, e_M\}$, retrieval-augmented fact-checking predicts a label $y \in \{\text{Support}, \text{Refute}, \text{Neutral}\}$ and, when required, an evidence-grounded rationale r, where *Neutral* denotes cases that are neither clearly supported nor clearly refuted by the available evidence, including evidence-insufficient, ambiguous, or conflicting cases depending on the original dataset annotation scheme. SEF-MAC verifies claims in two stages. First, a Structured Evidence Filtering Module $\mathcal{F}_{\text{SEF}}$ selects a refined subset:

$$\mathcal{E}' = \mathcal{F}_{\text{SEF}}(c, \mathcal{E}), \quad \mathcal{E}' \subseteq \mathcal{E}, \tag{1}$$

where $\mathcal{E}' = \{e_i \mid i \in I\}$, $I \subseteq \{1, \ldots, M\}$, and $|\mathcal{E}'| = m \leq M$. Then, a set of verifier agents ($\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$) collaboratively predicts the final output ($(y, r) = \Phi(\mathcal{A}, c, \mathcal{E}')$). In this paper, we set ($N = 3$). Below we detail the evidence filtering and challenge-and-validation procedures.

The overall pipeline of SEF-MAC is visualized in Figure 1. Our framework consists of two core components: the Structured Evidence Filtering Module for evidence denoising, and the Challenge-and-Validation Reasoning Approach for consensus convergence. The system first filters noisy retrieved evidence, then employs role-specialized verifier agents for deliberative fact-checking.

3.2 Structured Evidence Filtering Module

Given a claim c and its candidate evidence pool $E = \{e_1, \ldots, e_M\}$, we introduce Structured Evidence Filtering (SEF) to identify evidence that is factually useful for verification while reducing

interference from noisy or weakly related content. Rather than relying primarily on lexical or embedding similarity, SEF evaluates candidate evidence through structured factual consistency checks. SEF is training-free and operates before multi-agent reasoning.

Structured here means that domain knowledge is operationalized through the backbone LLM's parametric factual knowledge and structured evidence-grounded checks over entities, predicates, temporal expressions, numerical attributes, and domain-specific claim context, rather than through an external knowledge base. SEF decomposes the claim into atomic propositions as scoring aids, not separate aggregation variables, and adopts a compact, complementary decomposition of evidence quality into four dimensions: **core fact relevance** (s_r), **temporal consistency** (s_t), **entity alignment** (s_e), and **thematic coherence** (s_h), where $s_r, s_t, s_e, s_h \in [0, 1]$. These dimensions respectively capture whether the evidence addresses the claim's central fact, satisfies its temporal and entity constraints, and remains coherent with the overall claim context. The final consistency score is:

$$S_{\text{consist}}(c, e_i) = w_r s_r + w_t s_t + w_e s_e + w_h s_h, \tag{2}$$

where $w_r = 0.35$, $w_t = 0.20$, $w_e = 0.25$, $w_h = 0.20$. The ordering $w_r > w_e > w_t = w_h$ reflects this factual-utility prior, treating relevance and entity alignment as primary signals and temporal and thematic coherence as auxiliary signals, and is held fixed across all three benchmarks.

Scores follow a fixed rubric with anchors at 0, 0.5, and 1.0; intermediate values (0.25, 0.75) are allowed with a short justification. The four dimensions share the same 3-anchor structure summarized in Table 7 in Appendix A.2.2. The SEF output includes the evidence identifier, four sub-scores with justifications, and the weighted score; malformed numeric fields are clipped to $[0, 1]$ and missing fields default to 0.

To adapt to different evidence pools, we apply a dynamic threshold $\tau = \mu + \alpha\sigma$, computed per claim over its own candidate pool, where μ and σ are the mean and standard deviation of $\{S_{\text{consist}}(c, e_i)\}_{i=1}^{M}$. We set $\alpha = -0.2$ to bias SEF slightly toward retention over rejection: in noisy evidence pools, discarding borderline-relevant snippets risks removing the only verifiable signal, whereas including a weakly relevant snippet is less harmful than providing empty input to the reasoning agents. Snippets satisfying $S_{\text{consist}}(c, e_i) \geq \tau$ are retained. If no snippet passes the threshold, the highest-scoring one is kept to avoid empty reasoning input; ties are broken by original evidence order. This fallback does not prevent agents from predicting *Neutral* when retained evidence remains insufficient.

The refined set $\mathcal{E}' = \{e_i \mid i \in I\}$, $I \subseteq \{1, \ldots, M\}$, is forwarded to multi-agent verification without additional re-ranking.

3.3 Challenge-and-Validation Reasoning Approach

The approach has three stages: initial voting, dissent-driven deliberative debate, and consensus convergence. Given c and $\mathcal{E}'$, the three verifier agents first produce independent verdicts and rationales. Any non-unanimous judgment triggers lightweight debate to examine dissent and refine the decision.

The three agents share the same backbone LLM but use complementary role prompts. The **Evidence Sufficiency Agent** checks whether retained evidence is enough and avoids unsupported inference; the **Contradiction Detection Agent** searches for refuting evidence, entity/time mismatches, and inconsistent statements; and the **Logical Consistency Agent** checks the validity of the evidence-to-verdict reasoning chain, including overgeneralization, causal leaps, and missing steps. In the prompt, each role is asked to state not only its verdict but also the specific uncertainty or defect it inspected. These roles bias error-checking perspectives without forcing different labels, and the role prompts change only the inspection focus rather than the available evidence, label space, backbone model, or output format. All agents keep the same output schema: verdict, cited evidence identifiers, concise rationale, and inspected uncertainty or defect. Each agent receives c and identifier-tagged evidence snippets from $\mathcal{E}'$ only, without SEF scores; we pass at most the top-k retained snippets, with $k = 5$ in our experiments, and evidence order follows SEF retention order.

3.3.1 Initial Voting

Each agent independently outputs a preliminary verdict $\hat{y}_i \in \{\text{Support}, \text{Refute}, \text{Neutral}\}$, a structured rationale $\hat{r}_i$, and cited evidence identifiers. The rationale includes the judgment, evidence

Table 1: Class distribution of evaluation datasets.

Dataset	Support	Refute	Neutral	Total
CheckThat-T3 (Test)	717	2275	664	3656
CHEF (Test)	333	333	333	999
SciFact (Dev)	124	64	112	300

references, and core reasoning steps, yielding $Y = \{\hat{y}_1, \hat{y}_2, \hat{y}_3\}$ and $R = \{\hat{r}_1, \hat{r}_2, \hat{r}_3\}$. If all verdicts match, the approach enters consensus; otherwise, minority dissent or three-way disagreement triggers debate.

3.3.2 Dissent-Driven Deliberative Debate

When disagreement occurs, agents debate for at most two rounds. Dissenting agents challenge overlooked evidence, logical gaps, unsupported inference, or counterexamples; challenged agents respond with evidence-grounded defenses. Importantly, the rationales serve as explicit reasoning states: agents inspect the previous rationales to identify disputed reasoning steps and regenerate both verdicts and rationales using the debate history and $\mathcal{E}'$. The debate input contains previous verdicts, rationales, and cited evidence identifiers, with each agent providing a brief challenge or response before updating its prediction. Let $Y^{(t)}$ and $R^{(t)}$ denote outputs after round t, with $Y^{(0)} = Y$ and $R^{(0)} = R$ from initial voting:

$$(\hat{y}_i^{(t)}, \hat{r}_i^{(t)}) = \mathcal{D}\Big(a_i, Y^{(t-1)}, R^{(t-1)}, c, \mathcal{E}'\Big), t \in \{1, 2\}. \tag{3}$$

For brevity, cited evidence identifiers are omitted from the notation but are included in every agent output and debate input. A second round is used only if the first round does not reach unanimity.

3.3.3 Consensus Convergence and Final Verdict

After debate, agents output Y^* and R^*, where each rationale records the evidence and reasoning supporting its corresponding verdict. If unanimous, the shared verdict is used as y, together with a representative rationale from that unanimous round. Otherwise, rather than reverting to majority voting—which §5.5 shows is correct on only 40.58% of strict 2-vs-1 splits—we break ties by evidence grounding: we select the agent whose verdict cites the most retained snippets, returning its verdict and rationale as the final (y, r); ties prefer the Contradiction Detection Agent, then generation order for determinism. This grounds the arbitration in the SEF-filtered evidence rather than headcount, and by interrogating dissent rather than suppressing it, reduces majority bias on initially disagreed cases.

Implementation details, including decoding parameters, pseudocode, and the full parameter table, are provided in Appendix A.

4 Experimental Setup

4.1 Datasets

We evaluate SEF-MAC on three benchmark datasets: CheckThat-T3 for English numerical and temporal claims, CHEF for Chinese general-domain claims, and SciFact for English scientific claims. The class distributions are summarized in Table 1.

CheckThat-T3 (Alam et al., 2025). We evaluate on the English test set of CLEF 2025 CheckThat! Task 3, which focuses on verifying numerical and temporal claims. The test set contains 3,656 claims (True / False / Conflicting), with a skewed label distribution. For CheckThat-T3, inference is conducted using the original task labels: True, False, and Conflicting. We only report these labels under the normalized names Support, Refute, and Neutral in Table 1 and aggregate comparisons for consistency across datasets. Each claim is paired with up to five source documents retrieved via Google Search API, used as the candidate evidence pool under a high-noise web-retrieved evidence setting. This mapping is used only for unified three-way evaluation across datasets

Table 2: Main results on three fact-checking benchmarks.

Method	Model	CheckThat-T3		CHEF		SciFact	
		Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1
Closed-book Direct	gpt-4o-mini	46.83	45.45	49.75	49.58	50.00	44.00
	doubao-seed-2.0-lite	51.04	49.65	50.45	46.59	55.67	46.06
Closed-book CoT	gpt-4o-mini	49.89	46.94	43.14	40.79	34.33	25.99
	doubao-seed-2.0-lite	43.40	41.29	54.55	52.42	53.33	47.78
Standard RAG Direct	gpt-4o-mini	67.15	54.07	68.97	67.97	76.00	75.70
	doubao-seed-2.0-lite	63.54	54.52	68.97	68.64	77.00	77.08
Standard RAG CoT	gpt-4o-mini	65.67	56.24	57.46	56.31	73.00	72.34
	doubao-seed-2.0-lite	55.47	51.42	67.77	68.17	68.00	67.02
ReRead (Hu et al., 2023)	RoBERTa	–	–	71.24	69.52	–	–
PACAR (Zhao et al., 2024)	gpt-4o-mini	65.48	54.31	64.97	64.26	77.33	76.75
	doubao-seed-2.0-lite	63.76	53.74	68.17	67.58	79.00	79.56
D2D (Han et al., 2025)	gpt-4o-mini	63.27	53.95	64.16	63.46	76.33	75.92
	doubao-seed-2.0-lite	61.27	56.38	66.47	66.11	77.67	77.35
FIRE (Xie et al., 2025)	gpt-4o-mini	65.86	55.09	66.67	65.95	77.00	76.60
	doubao-seed-2.0-lite	62.39	52.37	68.47	68.02	78.67	78.09
SEF-MAC (Ours)	gpt-4o-mini	**69.20**	**56.59**	**73.17**	**72.52**	**79.33**	**79.48**
	doubao-seed-2.0-lite	**65.37**	**56.62**	**71.77**	**70.89**	**80.00**	**79.95**

CHEF (Hu et al., 2022). CHEF is a multi-domain Chinese fact-checking dataset. We use the 999-instance test split, where each claim is paired with up to five source documents retrieved via Google Search API as the candidate evidence pool. This setup evaluates the framework's ability to isolate sparse signals in high-noise scenarios.

SciFact (Wadden et al., 2020). We use the 300-instance development split under the gold evidence setting. Each claim is paired with annotator-associated abstracts where available, yielding 283 claims with associated evidence, representing a low-noise scenario for scientific domain reasoning. The remaining 17 claims without associated abstracts are retained in evaluation with an empty evidence block.

4.2 Implementation Details

We implement all agents of SEF-MAC with two lightweight backbones, `gpt-4o-mini-240718` and `doubao-seed-2.0-lite-260215`, in the main overall performance experiments (Section 5.1) to comprehensively evaluate model generality while balancing computational cost and performance. For subsequent ablation studies on the evidence filtering Module (Section 5.3) and reasoning approach (Section 5.4), we adopt only `gpt-4o-mini` as the unified backbone to eliminate model-induced confounding variables and ensure fair ablation comparison. We limit the generation length to 256 tokens to encourage concise, fact-based reasoning. This constraint standardizes response length during the multi-agent verification process.

4.3 Baselines

We evaluate SEF-MAC against a hierarchy of baselines to benchmark its performance.

Standard Inference Baselines. We implement four fundamental configurations by crossing knowledge sources with reasoning approach. Closed-book Direct/CoT baselines depend entirely on the LLM's internal parametric knowledge; the Direct mode applies direct prompting, whereas CoT (Chain-of-Thought) reasoning produces step-by-step logical inference (Wei et al., 2022). Standard RAG Direct/CoT corresponds to inference over the dataset-specific candidate evidence without our Structured Evidence Filtering Module. For CheckThat-T3 and CHEF, this evidence consists of raw retrieved source documents from the benchmark setting, whereas for SciFact it consists of the annotator-associated abstracts where available. The Direct mode produces a direct verdict, while

Table 3: Component ablation of SEF-MAC on three fact-checking benchmarks.

Method	CheckThat-T3		CHEF		SciFact	
	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1
SEF-MAC	**69.20**	**56.59**	**73.17**	**72.52**	**79.33**	**79.48**
w/o SEF	68.45	56.39	70.57	69.07	78.00	77.67
w/o C-V	67.53	54.49	71.07	69.55	76.67	76.26

the CoT mode uses chain-of-thought-style reasoning over the same unfiltered candidate evidence (Lewis et al., 2020).

Representative Frameworks. We compare our framework with representative domain-specific architectures. ReRead (Hu et al., 2023) is included as a published reference result for CHEF, since it is a representative evidence-faithful fact verification model evaluated on this dataset. PACAR (Zhao et al., 2024) is reproduced under the same backbone and evidence setting as SEF-MAC, serving as a representative planning-based LLM fact-checking baseline. Additionally, D2D (Han et al., 2025) reformulates misinformation detection as a structured multi-agent debate process to improve decision transparency, and FIRE performs fact-checking via an agent-based iterative loop that alternates between evidence retrieval and verification for more cost-effective reasoning.

Module-level Comparisons. Beyond system-level overall results, we conduct fine-grained evaluations of our designed sub-modules. The *Evidence Module* undergoes ablation studies and categorical source analysis to examine the impact of the Structured Evidence Filtering Module on evidence quality. The *Reasoning Strategies* are benchmarked against alternative multi-agent consensus strategies, such as **Self-Reflection** (Shinn et al., 2023), **Majority Voting**, and standard **Multi-agent Debate** (Du et al., 2024).

5 Results and Analysis

5.1 Main Results

Table 2 summarizes the comparative performance of our SEF-MAC framework against multiple baseline methods and SOTA approaches, evaluated with two different backbone models: `gpt-4o-mini` and `doubao-seed-2.0-lite`. To ensure a rigorous and fair evaluation, we report both Micro-F1 and Macro-F1 scores. Micro-F1 reflects overall prediction accuracy in the single-label setting, while Macro-F1 evaluates category-level balance across different claim labels. For CHEF, we cite the SOTA results from the original ReRead paper, which uses a RoBERTa backbone and thus serves as a reference rather than a like-for-like LLM comparison. All other multi-agent baselines (PACAR, D2D, and FIRE) on the three datasets are reproduced under both `gpt-4o-mini` and `doubao-seed-2.0-lite` backbones, consistent with SEF-MAC, to reduce performance bias caused by different base models.

The CoT Performance Gap. CoT generally underperforms Direct inference across the three benchmarks, particularly under noisy web-retrieval settings, with the exception of closed-book CHEF using the `doubao-seed-2.0-lite` backbone. This suggests that step-by-step reasoning can amplify interference from irrelevant retrieved evidence when no dedicated filtering stage is applied, whereas the smaller degradation on SciFact may reflect its cleaner, annotator-associated evidence setting.

Effectiveness of SEF-MAC. SEF-MAC delivers consistent gains on both Micro-F1 and Macro-F1 across all three benchmarks and both backbones. With `gpt-4o-mini`, SEF-MAC outperforms PACAR by 3.72 / 2.28 percentage points on CheckThat-T3, ReRead by 1.93 / 3.00 percentage points on CHEF, and our reproduced PACAR by 2.00 / 2.73 percentage points on SciFact in Micro-F1 and Macro-F1, respectively. Against the newer multi-agent baselines D2D and FIRE, SEF-MAC also leads on both metrics across all three datasets, with the largest Micro-F1 gain reaching 9.01 percentage points over D2D on CHEF. The `doubao-seed-2.0-lite` backbone exhibits the same pattern. Overall, these results suggest that SEF improves evidence quality while C-V provides a more targeted mechanism for resolving difficult or initially inconsistent cases.

Table 4: Performance before and after SEF filtering. *Mic*/*Mac* denote Micro-F1/Macro-F1 under Direct and CoT prompting.

Method	CheckThat-T3		CHEF		SciFact	
	Mic	Mac	Mic	Mac	Mic	Mac
Original (Direct)	67.15	54.07	68.97	67.97	76.00	75.70
Filtered (Direct)	**67.53**	54.49	**71.07**	**69.55**	**76.67**	**76.26**
Original (CoT)	65.67	56.24	57.46	56.31	73.00	72.34
Filtered (CoT)	66.17	**56.80**	58.56	57.74	74.33	73.68

Table 5: Evidence volume before and after SEF filtering. SEF prunes noisy snippets to yield a compact subset for multi-agent reasoning.

Metric	CheckThat-T3		CHEF		SciFact	
	Total	Avg.	Total	Avg.	Total	Avg.
Original	16259	4.45	4030	4.03	283	0.94
Filtered-out	6822	1.87	1532	1.53	14	0.05
Filtered	**9437**	**2.58**	**2498**	**2.50**	**269**	**0.89**

5.2 Component Ablation

To evaluate the complementary contributions of the two core modules, we compare the full SEF-MAC framework with two ablated variants: w/o SEF, which retains only the multi-agent reasoning component, and w/o C-V, which applies SEF followed by Direct inference.

As shown in Table 3, removing either module consistently reduces performance across all three benchmarks. Removing SEF causes the largest Micro-F1 drop on CHEF (2.60 points), while removing C-V has the largest effect on SciFact (2.66 points). These results indicate complementary strengths: SEF is particularly beneficial under noisy web retrieval, whereas C-V contributes more in settings that require additional reasoning and validation. Overall, combining SEF with C-V yields the strongest performance among the tested variants.

5.3 Analysis of Evidence Filtering

5.3.1 Impact of Evidence Filtering

To analyze the impact of SEF on fact-checking accuracy, we compare model performance before and after evidence filtering under both Direct and CoT reasoning settings. This analysis isolates the contribution of evidence refinement from the later multi-agent reasoning stage.

As shown in Table 4, SEF consistently improves both Direct and CoT reasoning across the three datasets. The improvement is largest on CHEF (up to 2.10 percentage points Micro-F1 under Direct inference), where web-retrieved evidence contains more irrelevant or weakly related snippets. By filtering evidence according to factual utility, temporal consistency, entity alignment, and thematic coherence, SEF reduces distraction from out-of-context information and provides a more stable input for downstream reasoning. On CheckThat-T3 and SciFact, the gains are more modest (0.38–1.33 percentage points Micro-F1): CheckThat-T3 claims depend on retrieving specific numerical and temporal facts rather than broad topical relevance, and SciFact's evidence pool is already constrained by annotator selection.

5.3.2 Evidence Volume Reduction

We further examine filtering efficiency by tracking the number of evidence snippets before and after SEF. This analysis measures whether SEF reduces evidence volume while preserving useful information for reasoning.

As summarized in Table 5, SEF substantially reduces the amount of evidence passed to the reasoning stage. On CheckThat-T3, it filters out approximately 42% of the retrieved snippets, reflecting the high-noise nature of open-domain web retrieval. On CHEF, it filters out approximately 38% of the retrieved snippets. On SciFact, only about 5% of evidence snippets are filtered out, consistent with the fact that the evidence pool is already constrained by the benchmark annotation process. The consistent performance gains under both high-noise and low-noise settings suggest that SEF is not merely a noise-removal heuristic, but a semantic refinement mechanism that helps concentrate reasoning on the most factually useful evidence. A complementary source-level analysis on CHEF shows that filtering shifts retained evidence toward authoritative sources; detailed results and source-classification criteria are provided in Appendix A.5.

Table 6: Comparison of different reasoning strategies on three benchmarks.

Method	CheckThat-T3		CHEF		SciFact	
	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1
Direct	67.53	54.49	71.07	69.55	76.67	76.26
CoT	66.17	56.80	58.56	57.74	74.33	73.68
Reflection	66.47	56.89	68.87	67.37	70.00	70.17
3-Agent Voting	66.58	57.01	72.47	71.11	76.00	75.61
Collab-Debate (3A)	66.93	52.35	72.37	71.06	72.00	71.38
Collab-Debate (6A)	65.86	52.13	70.97	69.52	71.33	70.49
Ours (C-V)	**69.20**	**56.59**	**73.17**	**72.52**	**79.33**	**79.48**

5.4 Ablation on Reasoning Strategies

After fixing the evidence filtering pipeline, we compare Challenge-and-Validation (C-V) with Direct/CoT inference, self-refinement (Madaan et al., 2023), three-agent majority voting, and 3-/6-agent collaborative debate (Du et al., 2024) under the same filtered-evidence setting. C-V uses three role-specialized agents and activates at most two debate rounds only when their initial verdicts disagree. As shown in Table 6, simply increasing the number of agents or adding rigid deliberation rounds does not consistently improve performance. For example, scaling collaborative debate from 3 agents to 6 agents reduces Micro-F1 by 1.07, 1.40, and 0.67 percentage points on CheckThat-T3, CHEF, and SciFact, respectively. This suggests that larger agent groups may introduce redundant or noisy opinions, and majority-driven consensus can dilute useful dissenting signals.

Compared with Direct inference, CoT, Reflection, majority voting, and collaborative debate, Challenge-and-Validation delivers the best Micro-F1 across all three benchmarks and the best Macro-F1 on CHEF and SciFact. Unlike standard voting or fixed collaborative debate, our method uses disagreement as a trigger for targeted validation. Therefore, deliberation is concentrated on difficult cases where agents initially disagree, rather than being uniformly applied to all samples. This design allows the framework to inspect logical flaws, evidence deficiencies, and contradictory viewpoints in peer opinions while avoiding unnecessary deliberation on already-consistent cases.

We further examine the effect of the debate-round budget and adaptive early stopping. The adaptive two-round setting provides the most consistent accuracy–cost trade-off across the three benchmarks; detailed results are provided in Appendix A.6.

5.5 Mechanism and Efficiency Analysis

Beyond aggregate performance, we analyze C-V's behavior on CheckThat-T3. Among the 3,656 claims, 16.77% (613 samples) trigger deliberation due to initial agent disagreement, while the rest achieve consensus without debate. For strict 2-vs-1 splits, naive majority voting obtains merely 40.58% accuracy. In contrast, C-V correctly resolves 57.10% of disagreed samples, and 81.89% of debated instances reach consensus after deliberation. Compared with the majority-voting baseline, C-V lifts Micro-F1 from 66.58% to 69.20% (+2.63pp, 95% bootstrap CI: [+1.70, +3.58], McNemar's exact test $p = 3.95 \times 10^{-8}$). Detailed role-level statistics are deferred to Appendix A.4.

6 Conclusion

We propose SEF-MAC, a unified multi-agent framework that couples Structured Evidence Filtering with Challenge-and-Validation Reasoning for retrieval-augmented fact-checking. SEF removes noisy evidence via structured factual consistency checks, while Challenge-and-Validation uses minority dissent to drive targeted deliberation and consensus convergence. Experiments on CheckThat-T3, CHEF, and SciFact demonstrate consistent Micro-F1 and Macro-F1 improvements over competitive retrieval-augmented and multi-agent baselines, and ablation studies confirm the complementary contribution of each module. For future work, we plan to reduce computational overhead through more efficient multi-agent scheduling and to explore extending the framework to multi-modal evidence settings.

AI USE STATEMENT

We used generative AI tools for assistance with language polishing, grammar checking, and improving the clarity of academic writing. We did not use generative AI tools to generate experimental results, conduct experiments, analyze data, or make scientific claims. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed SEF-MAC framework, including the evidence filtering mechanism and challenge-and-validation reasoning procedure, in the main paper. Experimental settings, benchmark configurations, and additional implementation details are provided in the appendix. We will release the source code and configuration files upon publication to facilitate reproduction of our results.

REFERENCES

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venktesh V. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, 2025. URL https://arxiv.org/abs/2503.14828.

Erik Arakelyan, Pasquale Minervini, Patrick Lewis, Pat Verga, and Isabelle Augenstein. Flare: Faithful logic-aided reasoning and exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23396–23414, 2025.

Jean-Flavien Bussotti and Paolo Papotti. Refining attention for explainable and noise-robust fact-checking with transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25487–25499, 2025.

Arbaaz Dharmavaram and Saqib Hakak. Sanctuary: An efficient evidence-based automated fact checking system. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 247–257, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Dzianis Piatrashyn, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072*, 2025.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms. *arXiv preprint arXiv:2505.17762*, 2025.

Chen Han, Wenzhen Zheng, and Xijin Tang. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15125–15140, 2025.

Bharatdeep Hazarika, Prasanna Devadiga, Pawan Kumar Rajpoot, Aditya U Baliga, Kishan Gurumurthy, Manish Jain, Manan Sharma, Ashish Shrivastava, Arya Suneesh, and Anshuman B Suresh. Tifin at checkthat! 2025: X-verify-multi-lingual nli-based fact checking with condensed evidence. In *CLEF (Working Notes)*, pp. 964–976, 2025.

Haorui He, Yupeng Li, Dacheng Wen, Yang Chen, Reynold Cheng, Donglong Chen, and Francis CM Lau. Debating truth: Debate-driven claim verification with multiple large language model agents. In *Proceedings of the ACM Web Conference 2026*, pp. 8851–8861, 2026.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 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*, pp. 3362–3376, 2022.

Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2319–2323, 2023.

Shaghayegh Kolli, Richard Rosenbaum, Timo Cavelius, Lasse Strothe, Andrii Lata, and Jana Diesner. Hybrid fact-checking that integrates knowledge graphs, large language models, and search-based retrieval agents improves interpretable claim verification. In Chen Zhang, Emily Allaway, Hua Shen, Lesly Miculicich, Yinqiao Li, Meryem M'hamdi, Peerat Limkonchotiwat, Richard He Bai, Santosh T.Y.S.S., Sophia Simeng Han, Surendrabikram Thapa, and Wiem Ben Rim (eds.), *Proceedings of the 9th Widening NLP Workshop*, pp. 106–115, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-351-7. doi: 10.18653/v1/2025.winlp-main.19. URL https://aclanthology.org/2025.winlp-main.19/.

Kumud Lakara, Georgia Channing, Christian Rupprecht, Juil Sock, Philip Torr, John Collomosse, and Christian Schroeder de Witt. Mad-sherlock: Multi-agent debate for visual misinformation detection. *arXiv preprint arXiv:2410.20140*, 2024.

Soohwan Lee and Kyungho Lee. Multi-agent consensus as a cognitive bias trigger in human-ai interaction. *arXiv preprint arXiv:2604.22277*, 2026.

Shuyi Liu, Yu-Ming Shang, and Xi Zhang. Truthfulrag: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32168–32176, 2026.

Chiyu Ma, Enpei Zhang, Yilun Zhao, Wenjun Liu, Yaning Jia, Peijun Qing, Lin Shi, Arman Cohan, Yujun Yan, and Soroush Vosoughi. Judging with many minds: Do more perspectives mean less prejudice? on bias amplifications and resistance in multi-agent based llm-as-judge. *arXiv preprint arXiv:2505.19477*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Shiyao Peng, Qianhe Zheng, Zhuodi Hao, Zichen Tang, Rongjin Li, Qing Huang, Jiayu Huang, Jiacheng Liu, Yifan Zhu, and Haihong E. Neocorrag: Less irrelevant information, more explicit evidence, and more effective recall via evidence chains. In *Proceedings of the ACM Web Conference 2026*, pp. 1899–1910, 2026.

Sascha Rolinger and Jin Liu. Graph-of-thoughts for fact-checking with large language models. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 266–273, 2025.

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. Face the facts! evaluating rag-based pipelines for professional fact-checking. In *Proceedings of the 18th International Natural Language Generation Conference*, pp. 846–865, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Satyam Shukla, Himanshu Dutta, and Pushpak Bhattacharyya. Recon, answer, verify: Agents in search of truth. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2429–2448, 2025.

Tingyu Song, Yilun Zhao, Siyue Zhang, Chen Zhao, and Arman Cohan. Limrank: Less is more for reasoning-intensive information reranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20636–20650, 2025.

Shengji Tang, Jianjian Cao, Weihao Lin, Jiale Hong, Bo Zhang, Shuyue Hu, Lei Bai, Tao Chen, Wanli Ouyang, and Peng Ye. Open-source llms collaboration beats closed-source llms: A scalable multi-agent system. *arXiv preprint arXiv:2507.14200*, 2025.

Max Upravitelev, Premtim Sahitaj, Arthur Hilbert, Veronika Solopova, Jing Yang, Nils Feldhus, Tatiana Anikina, Simon Ostermann, and Vera Schmitt. Exploring semantic filtering heuristics for efficient claim verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 229–237, 2025.

Dan Valeriu Voinea. Quantifying indeterminacy in news: A neutrosophic method for assessing fact-checking outcomes. *Neutrosophic Sets and Systems*, 90(1):13, 2025.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Xingchen Xiao, Heyan Huang, Runheng Liu, and Jincheng Xie. Mass-rag: Multi-agent synthesis retrieval-augmented generation. *arXiv preprint arXiv:2604.18509*, 2026.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. Fire: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2901–2914, 2025.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6002–6024, 2025.

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, 2024.

A Appendix

A.1 Full Inference Algorithm

We provide the full SEF-MAC inference pipeline, including Structured Evidence Filtering, selective multi-agent deliberation, and final arbitration.

This design concentrates additional reasoning budget on initially disagreed cases, while avoiding unnecessary deliberation on already-consistent examples.

Algorithm 1 SEF-MAC inference

Require: Claim c; candidate evidence pool $E = \{e_1, \ldots, e_M\}$; weights $(w_r, w_t, w_e, w_h) = (0.35, 0.20, 0.25, 0.20)$; retention bias $\alpha = -0.2$; verifier agents $N = 3$; max debate rounds $T = 2$; retained evidence cap $k = 5$

Ensure: Final verdict $y \in \{\texttt{Support}, \texttt{Refute}, \texttt{Neutral}\}$ and rationale r

for each $e_i \in E$ **do**
  Obtain SEF sub-scores $s_r, s_t, s_e, s_h \in [0, 1]$
  Compute $S_{\text{consist}}(c, e_i) = w_r s_r + w_t s_t + w_e s_e + w_h s_h$
end for
Compute claim-specific threshold $\tau = \mu + \alpha\sigma$, where μ, σ are the mean and standard deviation of $\{S_{\text{consist}}(c, e_i)\}_{i=1}^{M}$
Retain all snippets with $S_{\text{consist}}(c, e_i) \geq \tau$
if no snippet passes **then**
  Retain the highest-scoring snippet (tie broken by original evidence order)
end if
Let the retained set be E'
Pass at most top-k snippets in SEF retention order (identifier-tagged, without SEF scores) to three verifier agents: Evidence Sufficiency, Contradiction Detection, and Logical Consistency
Each agent independently predicts a verdict, cites evidence IDs, provides a concise rationale, and reports the inspected uncertainty or defect
if the three verdicts are unanimous **then**
  Return the shared verdict and the first generated unanimous rationale
end if
for $t = 1$ to T **do**
  Provide each agent with the previous verdicts, rationales, cited evidence IDs, and challenge/response history
  Each agent regenerates its verdict and rationale
  if unanimity is reached **then**
    Return the unanimous verdict and the first generated rationale in that unanimous round
  end if
end for
If disagreement remains, select the output citing the largest number of retained snippets; break ties by preferring the Contradiction Detection Agent and then generation order
Return the final verdict and rationale

A.2 Prompt Templates and Output Schemas

This section provides the prompt templates used in SEF-MAC. All verifier agents share the same backbone model within a given run and differ only in their role-specific inspection focus. Unless otherwise stated, outputs are required to follow a structured JSON-style schema for downstream parsing.

A.2.1 SEF Scoring Prompt

System prompt.

```
You are a Structured Evidence Filtering Module for fact-checking.
Use your parametric factual knowledge only to operationalize
structured checks over entities,
predicates, temporal expressions, numerical attributes, and
domain-specific claim context.
Do not perform final fact-check classification. Only score each
candidate evidence snippet for
whether it is useful evidence for verifying the claim.

Return valid JSON only. Do not wrap it in markdown.
```

User prompt template.

```
Claim:
{CLAIM}
```

```
Candidate evidence snippets:
{EVIDENCE_BLOCK}

Task:
1. Decompose the claim into at most 3 concise atomic propositions,
each under 12 words.
2. For each evidence snippet, score factual utility on four
dimensions:
   - s_r core fact relevance: 1.0 directly verifies the central
   proposition; 0.5 peripheral or incomplete; 0 irrelevant.
   - s_t temporal consistency: 1.0 temporally consistent; 0.5
   timing underspecified; 0 temporal contradiction.
   - s_e entity alignment: 1.0 identical entities or clear
   aliases; 0.5 partial matches; 0 different entities.
   - s_h thematic coherence: 1.0 same event or causal context; 0.5
   same broad topic without direct verification context; 0
   thematic disconnection.
   Intermediate values 0.25 and 0.75 are allowed when the snippet
   falls between anchors.
3. Provide very short justifications for each sub-score, each
under 8 words.

Return this exact JSON shape:
{
  "atomic_propositions": ["..."],
  "scores": [
    {
      "evidence_id": "E1",
      "s_r": 0.0,
      "s_t": 0.0,
      "s_e": 0.0,
      "s_h": 0.0,
      "justifications": {
        "s_r": "<=8 words",
        "s_t": "<=8 words",
        "s_e": "<=8 words",
        "s_h": "<=8 words"
      },
    }
  ]
}
```

Table 7: SEF scoring rubric.

Dimension	1.0	0.5	0
s_r Core fact relevance	Directly verifies central proposition	Peripheral / incomplete	Irrelevant
s_t Temporal consistency	Consistent timing	Underspecified timing	Temporal contradiction
s_e Entity alignment	Identical entities / aliases	Partial match	Different entities
s_h Thematic coherence	Same event / causal context	Same broad topic	Thematic disconnection

A.2.2 Common Verifier Output Schema

All three verifier agents use the following output schema:

```
{
```

```
  "verdict": "Support|Refute|Neutral",
  "cited_evidence": ["E1"],
  "rationale": "concise evidence-grounded rationale",
  "inspected_defect": "specific uncertainty, defect, or risk you
  inspected",
  "contradiction_analysis": "explicit contradiction or mismatch
  analysis, or 'none'"
}
```

A.2.3 Evidence Sufficiency Agent Prompt

System prompt.

```
You are the Evidence Sufficiency Agent in SEF-MAC, a multi-agent
fact-checking framework.

Role focus: Check whether the retained evidence is enough and
avoid unsupported inference. State the missing facts if evidence
is insufficient.

Use only the claim and identifier-tagged evidence snippets. The
evidence has already been selected by an upstream filtering
module; do not perform evidence filtering or ask for more
retrieval.

Return ONLY valid JSON:
{
  "verdict": "Support|Refute|Neutral",
  "cited_evidence": ["E1"],
  "rationale": "concise evidence-grounded rationale",
  "inspected_defect": "specific uncertainty, defect, or risk you
  inspected",
  "contradiction_analysis": "explicit contradiction or mismatch
  analysis, or 'none'"
}
```

User prompt template.

```
CLAIM:
{CLAIM}

EVIDENCE:
{EVIDENCE_BLOCK}

Produce your independent preliminary verdict, rationale, and cited
evidence identifiers.
```

A.2.4 Contradiction Detection Agent Prompt

System prompt.

```
You are the Contradiction Detection Agent in SEF-MAC, a
multi-agent fact-checking framework.

Role focus: Search for refuting evidence, entity/time mismatches,
numerical mismatches, and inconsistent statements.

Use only the claim and identifier-tagged evidence snippets. The
evidence has already been selected by an upstream filtering
module; do not perform evidence filtering or ask for more
retrieval.
```

```
Return ONLY valid JSON:
{
  "verdict": "Support|Refute|Neutral",
  "cited_evidence": ["E1"],
  "rationale": "concise evidence-grounded rationale",
  "inspected_defect": "specific uncertainty, defect, or risk you
  inspected",
  "contradiction_analysis": "explicit contradiction or mismatch
  analysis, or 'none'"
}
```

User prompt template.

```
CLAIM:
{CLAIM}

EVIDENCE:
{EVIDENCE_BLOCK}

Produce your independent preliminary verdict, rationale, and cited
evidence identifiers.
```

A.2.5 Logical Consistency Agent Prompt

System prompt.

```
You are the Logical Consistency Agent in SEF-MAC, a multi-agent
fact-checking framework.

Role focus: Check the evidence-to-verdict reasoning chain,
including overgeneralization, causal leaps, and missing steps.

Use only the claim and identifier-tagged evidence snippets. The
evidence has already been selected by an upstream filtering
module; do not perform evidence filtering or ask for more
retrieval.

Return ONLY valid JSON:
{
  "verdict": "Support|Refute|Neutral",
  "cited_evidence": ["E1"],
  "rationale": "concise evidence-grounded rationale",
  "inspected_defect": "specific uncertainty, defect, or risk you
  inspected",
  "contradiction_analysis": "explicit contradiction or mismatch
  analysis, or 'none'"
}
```

User prompt template.

```
CLAIM:
{CLAIM}

EVIDENCE:
{EVIDENCE_BLOCK}

Produce your independent preliminary verdict, rationale, and cited
evidence identifiers.
```

A.2.6 DEBATE PROMPT TEMPLATE

System prompt.

```
The debate stage uses the same role-specific SEF-MAC system prompt
as the
corresponding verifier agent. The role name and role focus are
instantiated as
Evidence Sufficiency Agent, Contradiction Detection Agent, or
Logical Consistency
Agent.
```

User prompt template.

```
CLAIM:
{CLAIM}

EVIDENCE:
{EVIDENCE_BLOCK}

PREVIOUS AGENT OUTPUTS:
{PREVIOUS_OUTPUTS}

DEBATE HISTORY:
{DEBATE_HISTORY}

This is deliberative debate round {ROUND_NO} of at most 2.
Challenge overlooked evidence, logical gaps, unsupported
inference, or counterexamples. If your previous verdict was
challenged, respond with an evidence-grounded defense. Then
regenerate your verdict and rationale.

Return ONLY valid JSON using the required schema.
```

A.2.7 DECODING AND PARSING NOTES

For reproducibility, this appendix additionally reports the decoding configuration used in prompt execution. Unless otherwise specified, verifier prompts are decoded with the default configuration in our evaluation script.

- Temperature: 0.1
- Top-p: not explicitly set; the provider default is used.
- Max tokens: 256
- Retry policy for malformed outputs: no extra LLM repair prompt is issued for malformed verifier outputs. The parser first searches for an explicit verdict field or verdict line, then falls back to keyword matching over the full response. Unparseable or insufficient outputs are treated conservatively as NEI.
- Number of retry attempts: API failures are retried up to 8 attempts in the evaluation runner, with linear backoff of 20 seconds multiplied by the attempt index. The Ark client also has an internal retry wrapper with 3 attempts and exponential backoff from a 2-second base delay.
- Whether system prompts are identical across datasets: yes. The same role-specific system prompt templates are used across datasets; only the user-side inputs, such as the claim text and retained evidence block, change.

A.3 ADDITIONAL IMPLEMENTATION DETAILS

For exact reproduction, all datasets are first converted into the same three-way claim-level format, where each instance contains a claim string and a normalized gold label in `Support`, `Refute`, or

Table 8: Core implementation settings of SEF-MAC.

Setting	Value
Backbone models	`gpt-4o-mini-240718`, `doubao-seed-2.0-lite-260215`
Verifier agents	3
SEF weights	$(0.35, 0.20, 0.25, 0.20)$
Retention bias	$\alpha = -0.2$
Retained evidence cap	$k = 5$
Max debate rounds	$T = 2$
Max generation length	256
Temperature	0.1
Top-p	Provider default; not explicitly set
Retry policy	API failures are retried up to 8 attempts with linear backoff; malformed verifier outputs are not repaired by an additional LLM call and are parsed conservatively, falling back to NEI when no valid verdict can be extracted.

Table 9: Initial verdict distribution of the three role-specialized agents on CheckThat-T3 before debate. Percentages are computed over all 3,656 test instances.

Role	Support	Refute	Neutral
Evidence Sufficiency	1,075 (29.40%)	1,846 (50.49%)	735 (20.10%)
Contradiction Detection	891 (24.37%)	2,384 (65.21%)	381 (10.42%)
Logical Consistency	1,092 (29.87%)	1,985 (54.29%)	579 (15.84%)

`Neutral`. For datasets with original labels outside this space, labels are mapped before evaluation and the model is always evaluated on the normalized three-way labels.

Evidence retrieval follows the same Google-based pipeline for the main evidence-augmented setting. For each claim on CheckThat-T3 and CHEF, we issue the query $\{$`claim`$\}$ to the Google Search API and retain at most the top five returned organic results. Each retrieved item is stored with its rank, title, URL, and snippet. Titles are truncated to 200 characters, URLs to 500 characters, and snippets to 700 characters. Retrieval results are written incrementally to a resumable JSONL cache and then consolidated into a final JSON cache. During evaluation, the cached Google evidence is loaded by claim index or claim text, and only the top-k evidence snippets are passed forward.

SEF is applied after the Google Search API retrieval stage and before multi-agent reasoning. Verifier agents do not observe SEF scores and only receive the retained, identifier-tagged evidence snippets. The evaluation script reports accuracy, Micro-F1, Macro-F1, and per-class precision, recall, F1, support, and prediction counts. Since fact-checking here is a single-label three-way classification task, Micro-F1 is numerically identical to accuracy. The evaluation loop does not shuffle examples in the main full-dataset runs, and the sampling seed is fixed to 42 only for stratified debugging runs.

A.4 Role-Level Initial Disagreement Analysis

To further examine whether the three role-specialized agents exhibit different initial behaviors, we analyze their preliminary verdict distributions on CheckThat-T3 before any debate is triggered. Table 9 shows that the three roles do not produce identical label distributions. The Contradiction Detection Agent predicts Refute more frequently than the other two agents, while the Evidence Sufficiency Agent assigns Neutral more often. This pattern is consistent with their role prompts: the contradiction-focused agent is more sensitive to refuting or mismatched evidence, whereas the sufficiency-focused agent is more conservative when evidence is incomplete.

We also analyze which role becomes the minority dissenter in initially disagreed cases. The analysis is restricted to strict 2-vs-1 disagreements; 3-way ties, where all three agents predict different labels, are excluded from the minority dissenter count. As shown in Table 10, among 584 strict 2-vs-1 disagreements, the Contradiction Detection Agent accounts for the largest share of minority dissent. This suggests that contradiction-oriented inspection often introduces a distinct opposing view rather than merely replicating the majority judgment.

Table 10: Minority dissenter distribution on CheckThat-T3. The count is computed only over 584 initial strict 2-vs-1 disagreement cases; 29 initial 3-way ties are excluded.

Role	Minority dissenter count	Share among 2-vs-1 disagreements
Evidence Sufficiency	187	32.02%
Contradiction Detection	360	61.64%
Logical Consistency	37	6.34%

Figure 2 further summarizes the sample-level flow of Challenge-and-Validation on CheckThat-T3. Most samples reach initial consensus and therefore do not trigger debate, while C-V focuses computation on the 613 initially disagreed cases. Among these debated cases, 502 reach consensus after debate, whereas 111 remain unresolved and are handled by the fallback arbitration rule.

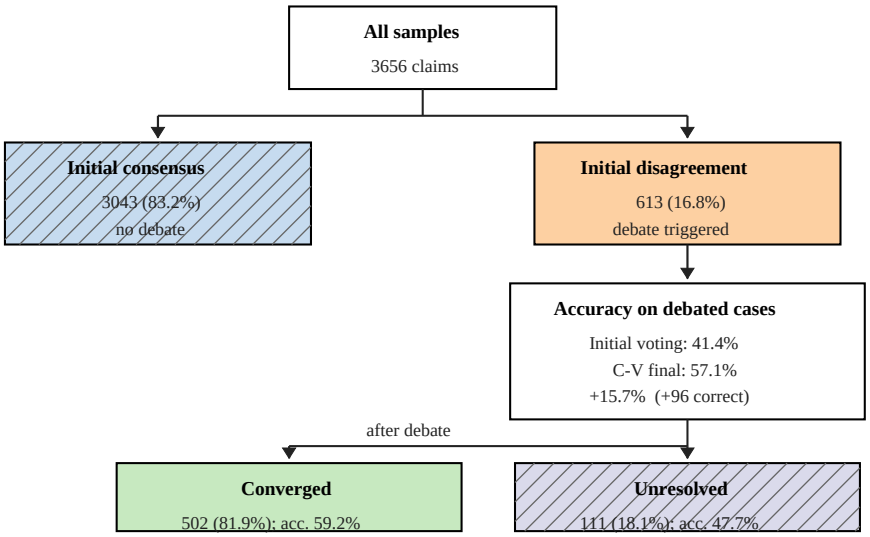

Figure 2: Sample flow of Challenge-and-Validation on CheckThat-T3. Among 3,656 claims, 3,043 cases reach initial consensus and do not trigger debate, while 613 initially disagreed cases enter the debate stage. After debate, 502 cases converge to consensus and 111 remain unresolved. Accuracy values are reported for the corresponding subsets.

A.5 Source Distribution Analysis

To complement the evidence-volume analysis, we examine how SEF changes the source composition of retained evidence. We conduct this analysis on CHEF, whose open-domain web retrieval produces heterogeneous source types. CheckThat-T3 follows a similar web-retrieval setting, whereas SciFact uses annotator-associated scientific abstracts and therefore exhibits substantially less source diversity.

Figure 3 illustrates the shift in evidence composition on CHEF, where categories are assigned by a deterministic URL-domain classifier (Appendix A.5). After filtering, the proportion of Authoritative Media increases by 2.73 percentage points, while the proportions of Commercial and Other sources decrease by 2.67 and 4.31 percentage points, respectively. This pattern suggests that semantically consistent and directly verifiable sources are more likely to be retained. In complex or evidence-insufficient cases, the absence of relevant information in authoritative sources can also serve as a useful signal for predicting insufficient evidence, reducing the risk of forced incorrect judgments.

For the source-distribution analysis in Appendix A.5.1, we use a deterministic, rule-based classifier that maps each evidence snippet to a source category solely from its URL domain. The rules below are applied in order as regular-expression substring matches; the first rule that matches an evidence URL assigns its category, and any URL matching no rule is assigned to *Other*. This avoids any LLM-

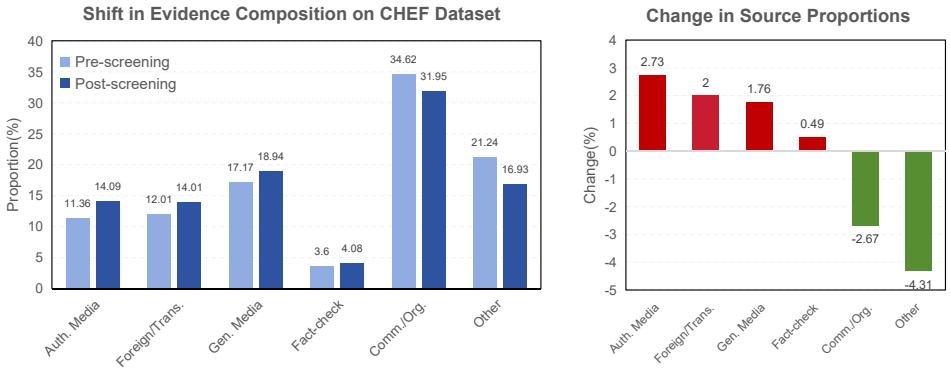

Figure 3: Shift in evidence-source composition on CHEF before and after Structured Evidence Filtering (SEF). Left: source-category proportions. Right: percentage-point changes after filtering.

based or subjective judgement and makes the reported percentages fully reproducible. Snippets without a URL are likewise assigned to *Other*.

Table 11: Source categories used in the CHEF distribution analysis. Domain patterns are matched as substrings against the evidence URL.

Source category	Matching URL domains
Authoritative Media	xinhuanet.com, news.cn, cctv.com, people.com.cn, chinanews.com.cn, *.gov.cn, 12377.cn
Commercial / General Media	sohu.com, qq.com, 163.com, ifeng.com, sina.com.cn, weibo.com
Knowledge / Community	zhihu.com, baike.baidu.com
Fact-checking Platform	cofacts.tw
Overseas / Translated	translate.google.com, yahoo.com
Other	any remaining or unmatched domain (catch-all)

Under this scheme, *Authoritative Media* is restricted to state-run news agencies, party-affiliated press, and government portals (e.g., Xinhua, *CCTV*, *People's Daily*, and ".gov.cn" regulatory sites). Commercial web portals and social platforms are deliberately *not* treated as authoritative, even when they carry reporting, and self-published blogs, forums, and unmatched domains fall into *Other*. The Authoritative Media gain reported in Figure 3 therefore reflects a shift toward verifiable, institutionally published evidence after SEF filtering, rather than a re-weighting of commercial or user-generated content.

A.6 Round-Budget Ablation of Challenge-and-Validation

To assess the necessity of adaptive debate depth, we compare three Challenge-and-Validation variants using the same `gpt-4o-mini` source traces and the same evidence-grounded fallback arbitration rule. **Adaptive-1** allows at most one debate round, **Adaptive-2** is our default setting that stops once consensus is reached with at most two rounds, and **Fixed-2** forces all initially disagreed cases to complete two rounds. The Total calls column counts one SEF scoring call plus the conditional C-V calls per claim.

Increasing the adaptive budget from one to two rounds consistently improves Micro-F1, with the clearest gain on SciFact, where Adaptive-2 improves Micro-F1 from 78.33% to 79.33% and Macro-F1 from 77.67% to 79.48%. This suggests that a second round remains useful for a small subset of unresolved cases. In contrast, forcing two rounds for all initially disagreed cases yields no consistent benefit. On CheckThat-T3, Fixed-2 improves only marginally over Adaptive-2 (69.20% $\rightarrow$ 69.28% Micro-F1) at a higher cost (4.67 $\rightarrow$ 5.01 total calls). On CHEF, Fixed-2 degrades both Micro-F1 and Macro-F1. On SciFact, it matches Adaptive-2 in Micro-F1 but lowers Macro-F1 while still requiring more total calls. Overall, these results support adaptive early stopping with a maximum

Table 12: Round-budget ablation of Challenge-and-Validation across three benchmarks.

Dataset	Method	Mic-F1	Mac-F1	Total Calls
CheckThat-T3	Adaptive-1	69.06	56.42	4.50
	Adaptive-2	69.20	56.59	4.67
	Fixed-2	69.28	56.64	5.01
CHEF	Adaptive-1	73.07	73.00	4.56
	Adaptive-2	73.17	72.52	4.66
	Fixed-2	72.07	71.99	5.12
SciFact	Adaptive-1	78.33	77.67	4.42
	Adaptive-2	79.33	79.48	4.46
	Fixed-2	79.33	78.67	4.84

of two rounds: it captures most of the benefit of additional deliberation without uniformly inflating total calls or per-claim latency, as debate fires only on the small initially-disagreed subset.

A.7 Statistical Testing Details

We assess statistical reliability by comparing Challenge-and-Validation against the initial majority-voting baseline with deterministic tie-breaking on the same 3,656 test instances. The baseline uses the three initial verifier outputs only; when the three agents produce a three-way tie, a fixed deterministic tie-breaking rule is applied.

For Accuracy/Micro-F1 and Macro-F1, we use unstratified paired bootstrap resampling at the claim level with 10,000 resamples and random seed 20260810. No multiple-comparison correction is applied because Accuracy/Micro-F1 is the primary significance target and Macro-F1 is reported as a secondary diagnostic metric.

Challenge-and-Validation improves Accuracy/Micro-F1 from 66.58% to 69.20%, corresponding to a paired bootstrap difference of +2.63 percentage points with a 95% confidence interval of [+1.70, +3.58].

We additionally apply McNemar's exact two-sided test at the sample level. The test yields $p = 3.95 \times 10^{-8}$, with 104 cases where the baseline is correct but Challenge-and-Validation is wrong and 200 cases where the baseline is wrong but Challenge-and-Validation is correct. These results support a statistically significant gain in overall Accuracy/Micro-F1.

A.8 SEF Weight Sensitivity Study

The consistency score in Eq. equation 2 combines the four SEF dimensions with fixed weights $(w_r, w_t, w_e, w_h) = (0.35, 0.20, 0.25, 0.20)$, set by a factual-utility prior and held constant across all three benchmarks. To verify that the reported conclusions do not depend on the exact choice of these values, we rerun SEF-MAC under an equal-weight variant that replaces them with $(0.25, 0.25, 0.25, 0.25)$ while keeping every other component—dynamic threshold, retention fallback, C-V debate, and the backbone model—identical. Table 13 reports the results on CHEF, the high-noise web-retrieval benchmark on which SEF has the largest effect.

Table 13: Sensitivity to SEF weighting on CHEF (Adaptive-2, `gpt-4o-mini`).

SEF weighting	Micro-F1	Macro-F1
Factual-utility $(0.35, 0.20, 0.25, 0.20)$	73.17	72.52
Equal $(0.25, 0.25, 0.25, 0.25)$	73.07	72.48

Replacing the factual-utility weights with equal ones changes Micro-F1 by *0.10* percentage points and Macro-F1 by *0.04* percentage points, which indicates that the performance reported in the main text is not sensitive to the precise weighting: the gain from SEF persists when all four dimensions are treated equally, and the factual-utility prior only re-ranks evidence at the margin.

A.9 CASE STUDY

To illustrate how SEF-MAC operates end-to-end, we present a detailed trace from CheckThat-T3. The case traces a claim from candidate evidence through SEF filtering, initial verdicts, dissent-driven debate, and final consensus. The shared backbone is `gpt-4o-mini`. This case is selected because it involves a subtle semantic precision error: a legally weaker statement about "plausible" rights and provisional measures is incorrectly upgraded into a definitive ruling that genocide was being committed.

A.9.1 CHECKTHAT-T3 CASE: SEMANTIC PRECISION DISTORTION

This case illustrates how SEF filters temporally misaligned and overly broad evidence, and how the Challenge-and-Validation mechanism corrects an initial majority misjudgment caused by a subtle semantic distinction between "plausible" and "ruled". The example is representative of claims where the surface topic is correct, but the asserted legal or factual strength is distorted.

Claim.

> On January 26, 2024, the International Court of Justice ruled Israel is committing genocide against Palestinians in Gaza.

Case metadata. The case index is `3241`. The gold label is *Refute*; SEF-MAC predicts *Refute*. The final selection reason is unanimous consensus after debate.

Candidate evidence. The Google Search API returned five candidate evidence snippets:

1. **Israel-Gaza: What did the ICJ ruling really say?** (BBC).
   `https://www.bbc.com/news/articles/c3g9g63jl17o`
   Snippet: The purpose of the ruling was to declare that South Africa had a right to bring its case against Israel and that Palestinians had "plausible rights to ..."
2. **ICJ says it's "plausible" Israel committed genocide in Gaza** (NPR).
   `https://www.npr.org/2024/01/26/1227078791/icj-israel-genocide-gaza-palestinians-south-africa`
   Snippet: The International Court of Justice ordered Israel to ensure its military does not take actions that violate the Genocide Convention, but the court did not make a final determination that genocide had occurred.
3. **Israel Not Complying with World Court Order in Genocide Case** (Human Rights Watch).
   Human Rights Watch, Feb. 26, 2024
   Snippet: The Israeli government has failed to comply with at least one measure in the legally binding order from the International Court of Justice ...
4. **Summary of the Order of 26 January 2024** (ICJ official).
   `https://www.icj-cij.org/node/203454`
   Snippet: The Court indicated provisional measures concerning the prevention and punishment of the crime of genocide in Gaza, while assessing plausible rights at this stage rather than deciding the merits of whether genocide was being committed.
5. **Palestinian genocide allegations** (Wikipedia).
   `https://en.wikipedia.org/wiki/Palestinian_genocide_allegations`
   Snippet: Since its foundation in 1948, Israel has been accused of carrying out genocide against Palestinians during the Israeli–Palestinian conflict.

SEF atomic propositions. SEF decomposes the claim into three atomic propositions as scoring aids: (1) the ICJ issued a ruling on January 26, 2024; (2) the ruling concerns Israel and Palestinians in Gaza; and (3) the ruling finds that Israel committed genocide.

Table 14: SEF scoring results for the CheckThat-T3 case.

ID	Source	s_r	s_t	s_e	s_h	Score
E1	BBC	0.50	0.50	1.00	1.00	0.725
E2	NPR	0.75	1.00	1.00	1.00	0.913
E3	HRW	0.50	0.00	1.00	0.50	0.525
E4	ICJ	1.00	1.00	1.00	1.00	1.000
E5	Wiki	0.25	0.00	0.50	0.50	0.313

Table 15: Verdict changes during Challenge-and-Validation for the case study.

Agent	Round 0	Key inspected issue	Round 1
Evidence Sufficiency	Neutral	No definitive merits ruling	Refute
Contradiction Detection	Refute	"Plausible" does not entail "ruled"	Refute
Logical Consistency	Neutral	Overgeneralization from provisional measures	Refute

SEF scoring and filtering. Table 14 reports the four SEF sub-scores and the weighted consistency score $S_{\text{consist}} = 0.35s_r + 0.20s_t + 0.25s_e + 0.20s_h$. With the dynamic threshold $\tau = \mu + \alpha\sigma$ and $\alpha = -0.2$, the threshold for this claim is approximately 0.55. Thus, E1, E2, and E4 are retained, while E3 and E5 are discarded.

Filtering rationale. E1 is retained because it covers the ICJ Gaza ruling, although its date and ruling details are only partially specified. E2 is retained because it directly covers the January 26, 2024 ICJ event and explicitly frames the legal finding as "plausible" rather than final. E4 is retained because it is the ICJ primary legal document and fully matches the target ruling context, including the distinction between provisional measures and a merits-stage determination. E3 is discarded because it discusses post-ruling compliance on February 26, 2024 rather than the January 26 ruling itself, creating a temporal mismatch. E5 is discarded because it is a broad historical encyclopedia entry about genocide allegations and lacks the specific 2024 legal-event context.

The retained set $\{E1, E2, E4\}$ is sufficient for the downstream agents because it preserves the relevant date, institution, parties, and legal context, while removing evidence that is temporally shifted or too general to verify the exact claim.

Initial voting. The retained evidence set $\{E1, E2, E4\}$ is then passed to the three verifier agents. In Round 0, the Evidence Sufficiency Agent predicts *Neutral*, reasoning that the evidence discusses plausibility and procedural orders but not a definitive ruling. The Contradiction Detection Agent predicts *Refute*, identifying that "plausible" does not entail "ruled". The Logical Consistency Agent predicts *Neutral*, noting that the claim overgeneralizes the court's plausibility assessment. This produces an initial 2-vs-1 split (Neutral, Refute, Neutral), which would lead simple majority voting to an incorrect *Neutral* verdict.

Debate and convergence. During Round 1, the Contradiction Detection Agent challenges the two *Neutral* verdicts by emphasizing the semantic distinction between the court's actual language of plausibility and the claim's stronger assertion that the court definitively ruled Israel was committing genocide. After re-examining E2 and E4, both the Evidence Sufficiency Agent and the Logical Consistency Agent update their verdicts to *Refute*. The three agents therefore converge to a unanimous *Refute* verdict.

Final verdict and insight. The final verdict is **Refute**, matching the gold label. The key reason is that the claim does not merely state that the ICJ considered genocide-related allegations or issued provisional measures. Instead, it asserts that the ICJ ruled Israel was committing genocide against Palestinians in Gaza. The retained evidence shows a different legal status: the ICJ recognized plausible rights and ordered provisional measures, but it did not make a definitive merits-stage determination that Israel was committing genocide. Therefore, the claim is contradicted by the legal meaning of the ruling, rather than being only evidence-insufficient.

This case shows that SEF can remove temporally misaligned or overly broad evidence before reasoning, while Challenge-and-Validation can preserve useful minority dissent. In particular, the

contradiction-focused agent captures the key semantic error and prevents the initial majority *Neutral* judgment from suppressing a more precise *Refute* interpretation. This kind of semantic precision distortion is harder than outright fabrication and is exactly where dissent-driven validation is useful.