

Perceive-Then-Adjudicate: A Multi-Agent Framework for Resolving Knowledge-Evidence Conflicts in LLM Fact-Checking

Yongcheng Zhang¹, Xin Zhang², Fei Li^{1*}, Chong Teng¹,
Jiangming Yang², Donghong Ji^{1*}, Zhuang Li³

¹ School of Cyber Science and Engineering, Wuhan University, China

²Ant International, ³ School of Computing Technologies, RMIT University, Australia

{zhangyongcheng, lifei_csnlp, tengchong, dhji}@whu.edu.cn

{evan.zx, jmyang}@ant-intl.com, zhuang.li@rmit.edu.au

Abstract

Fact-checking has transitioned from traditional semantic matching toward active reasoning powered by Large Language Models (LLMs), manifesting as the interplay between internal parametric knowledge and external grounded evidence. However, this paradigm introduces two critical challenges: (i) excessive reliance on external evidence even when it conflicts with correct internal judgment, and (ii) uncritical acceptance of a consensus even when both sources are erroneous. To address these limitations, we propose a Multi-Agent Arbitration Framework where a meta-cognitive Adjudication Agent (AA) orchestrates a Knowledge Agent (KA) and an Evidence Agent (EA). Through hybrid reinforcement and imitation learning, the AA perceives conflict signals between KA and EA to adaptively calibrate judgmental autonomy against evidentiary support. Upon detecting a potentially erroneous consensus, the AA invokes a Reflective Rejection mechanism to scrutinize and disrupt deceptive alignment. Empirical evaluations on Snopes and PolitiFact demonstrate that our framework consistently outperforms state-of-the-art baselines. Notably, it effectively mitigates over-reliance errors while maintaining high fidelity in correct-consensus scenarios, recovering 26.2% of dual-error cases where static systems typically succumb to superficial consistency. Our code is publicly available at <https://github.com/Zhang-Gorden/Perceive-Then-Adjudicate>.

1 Introduction

Automatic fact-checking aims to verify the veracity of claims by evaluating their consistency with evidence, serving as a pivotal defense against the proliferation of misinformation (Lin et al., 2025; Mujahid et al., 2025). Traditional methods have framed this task as a semantic matching paradigm,

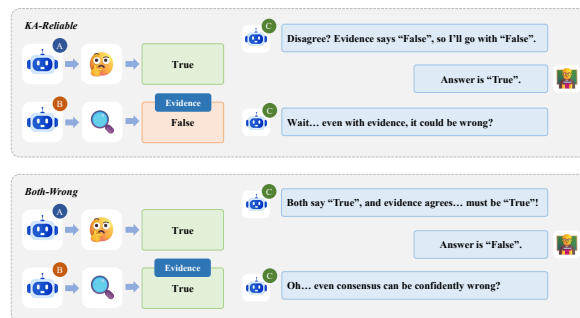


Figure 1: Illustrations of systemic failures in LLM-based fact-checking. (a) *KA-Reliable*: The model erroneously overturns its correct internal judgment due to excessive reliance on external evidence. (b) *Both-Wrong*: The model uncritically trusts an erroneous consensus between its internal knowledge and external evidence.

relying on explicit alignment mechanisms to bridge claims and retrieved evidence (Zhang et al., 2024b; Chen et al., 2025). With the rapid advancement of Large Language Models (LLMs), this paradigm has undergone a fundamental shift: LLMs are no longer passive evidence matchers but act as active reasoning agents, where fact-checking manifests as a sophisticated interplay between their inherent parametric knowledge and external evidentiary support (Wang et al., 2025; Jiang et al., 2025).

However, this shift uncovers a subtle yet critical form of cognitive bias (Zhang et al., 2024a; Guo et al., 2024). For example, Zhu et al. (2025) demonstrated that in question-answering settings, LLMs exhibit a strong tendency toward contextual over-reliance, overriding correct internal priors when confronted with a fabricated consensus of incorrect options. While such findings often stem from artificially induced false consensus (Sharma et al., 2024; Khan et al., 2024), fact-checking pipelines typically leverage high-quality, human-annotated “gold” evidence from authoritative sources (Vlachos and Riedel, 2014; Popat et al., 2017). This raises a fundamental question: **does reliable evidence al-**

*Corresponding authors

ways facilitate correct reasoning? We hypothesize that external evidence intended to ground judgments may paradoxically erode the model’s judgmental autonomy. To test this hypothesis, we conduct a controlled study comparing two variants of the same LLM: a **Knowledge Agent (KA)**, relying solely on its parametric knowledge, and an **Evidence Agent (EA)**, augmented with gold evidence. Our results reveal that while evidence yields a 12.3% accuracy gain, this improvement masks a counter-intuitive vulnerability: in 9.1% of cases, the incorporation of evidence causes the model to overturn a previously correct internal judgment.

To characterize these interactions, we categorize instances into a taxonomy based on agent agreement and veracity: *Both-Correct*, *KA-Reliable*, *EA-Reliable*, and *Both-Wrong*. Notably, in *Both-Wrong* cases, the system exhibits pronounced **reasoning inertia**, passively conforming to an erroneous consensus without proactive questioning or self-correction (Egle et al., 2022; Perez et al., 2023; Dong et al., 2024). These patterns reveal a deeper dilemma in LLM-based fact-checking (Figure 1): when agents disagree, the model often misinterprets the neutral instruction of “refer to evidence” as an **implicit mandate** of “prioritize evidence” (Lee and Souther, 2020; Askell et al., 2021), thereby marginalizing analytical reasoning in favor of superficial claim-evidence matching. Conversely, when agents agree, the model rarely scrutinizes the outcomes, uncritically treating consensus as a sufficient proxy for ground truth, regardless of its underlying factual accuracy. This poses a critical question: **How can LLMs reliably arbitrate between internal knowledge and external evidence when both conflicting signals and erroneous consensus lead to systemic failures?**

To address these challenges, we propose a **Multi-Agent Arbitration Framework** that orchestrates the interplay between internal judgment and external evidence. Our framework coordinates three specialized components: a **KA**, an **EA**, and a meta-cognitive **Adjudication Agent (AA)** (Kadavath et al., 2022; Madaan et al., 2023). While the KA and EA generate initial verdicts, the AA perceives their conflict and consensus signals to govern the final fact-checking decision. In cases of conflicting signals, the AA calibrates autonomous reasoning against evidentiary support; more crucially, upon detecting a potentially erroneous consensus, it invokes a **Reflective Rejection** mechanism to scrutinize and break deceptive alignment.

Trained via hybrid reinforcement and imitation learning (Liu et al., 2022), our framework adopts a **“Perceive-Then-Adjudicate”** policy. Empirical results demonstrate that this design effectively mitigates over-reliance errors and equips the system with self-calibrating capabilities, enabling robust integration of external information without sacrificing judgmental autonomy. Our contributions are:

- **Empirical Characterization and Taxonomy:** We identify and quantify two critical challenges in LLM-based fact-checking: **implicit mandate** and **reasoning inertia**. We formalize these via a taxonomy categorizing agent interactions based on the alignment and veracity of internal versus external signals.
- **Multi-Agent Arbitration Framework:** We propose a framework where an Adjudication Agent (AA) orchestrates the Knowledge Agent (KA) and Evidence Agent (EA). It calibrates judgmental autonomy against evidentiary support, employing **Reflective Rejection** to scrutinize and break deceptive alignment.
- **Hybrid Learning and Evaluation:** We combine reinforcement and imitation learning to optimize a **“Perceive-Then-Adjudicate”** policy. Evaluations on Snopes and PolitiFact show our framework outperforms SOTA baselines, particularly in resolving knowledge-evidence conflicts and *“Both-Wrong”* failures.

2 Pilot Study: Quantifying Biases

To empirically investigate the hypothesized interplay between internal judgment and external evidence, we conduct a preliminary analysis using two representative LLMs, Qwen2.5-7B-Instruct (Qwen) (Yang et al., 2024) and Llama3.1-8B-Instruct (Llama) (Team, 2024), on the training sets of Snopes and PolitiFact (Vlachos and Riedel, 2014; Popat et al., 2017) (detailed in Section 4.1). Notably, this analysis employs zero-shot inference wherein LLMs have no access to labels. Our analysis reveals the following three phenomena:

The Double-Edged Nature of Evidence. As shown in Table 1, gold evidence consistently boosts global performance gains. Average accuracy increases by 12% and 12.5% for Qwen and Llama, respectively. Nevertheless, these aggregate improvements mask critical local failures.

Models	Agents	Snopes		PolitiFact	
		Acc	F1	Acc	F1
Qwen	KA	0.73	0.66	0.50	0.37
	EA	0.82	0.83	0.65	0.64
Llama	KA	0.72	0.69	0.52	0.43
	EA	0.83	0.82	0.66	0.65

Table 1: Performance comparison between KA and EA. The superior Accuracy (Acc) and F1-score (F1) of EA across datasets validate the utility of external evidence, while providing an empirical basis for investigating potential over-reliance on external signals.

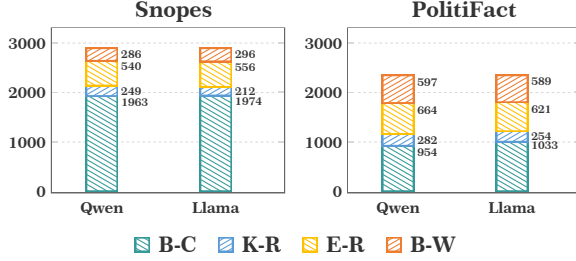


Figure 2: Distribution of agent interaction types. We categorize instances into: *Both-Correct* (B-C), *KA-Reliable* (K-R), *EA-Reliable* (E-R), and *Both-Wrong* (B-W). The proportions of K-R and B-W cases empirically confirm the existence of implicit mandate (correct internal judgment overridden by external evidence) and reasoning inertia (failure under erroneous consensus), respectively.

Implicit Mandate and Reasoning Inertia. We categorize agent interactions into a taxonomy of four types: (i) *Both-Correct*, (ii) *KA-Reliable*, (iii) *EA-Reliable*, and (iv) *Both-Wrong*. Figure 2 shows that while consensus occupies the majority of cases, underlying error patterns reveal deep-seated systemic biases. Crucially, in 9.1% of *KA-Reliable* instances, evidence integration paradoxically overrides correct internal priors, validating the implicit mandate to prioritize external signals over internal judgment. Furthermore, Figure 2 highlights reasoning inertia in *Both-Wrong* samples (15.9% of total errors), where both agents converge on erroneous consensus without triggering proactive scrutiny.

Strategy Rigidity. We further investigate whether agents adaptively adjust their reasoning strategies based on task difficulty. As shown in Figure 3, we observe a pervasive strategy rigidity, where agents default to low-cost heuristics regardless of claim complexity. Specifically, the KA consistently favors Fast (direct inference) over Step (chain-of-thought) in 54.9% of cases, while the EA relies on superficial Sum (summary-level comparison) rather than fine-grained Piece-wise

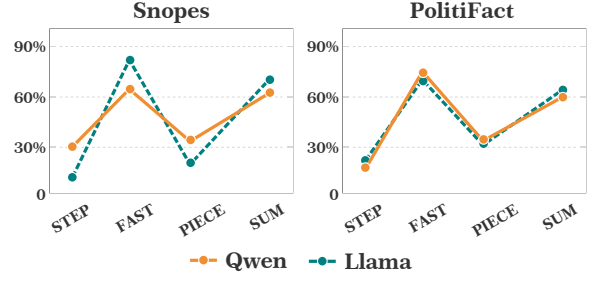


Figure 3: Distribution of default strategies. The dominance of low-cost patterns (Fast and Sum) reveals strategy rigidity, where agents fail to autonomously escalate to higher-order, fine-grained reasoning (e.g., Step or Piece-wise) when confronted with complex claims.

matching in 33.5% of instances (see Figure 4 for definitions of these four strategies). This rigidity reinforces reasoning inertia, especially in *Both-Wrong* cases where agents succumb to the erroneous consensus instead of escalating to deliberative reasoning to uncover latent conflicts.

3 Methodology

3.1 Overview

Guided by the insights from Section 2, we propose a **Multi-Agent Arbitration Framework** governed by a “**Perceive-Then-Adjudicate**” policy. As illustrated in Figure 4, the framework coordinates three specialized agents: **Knowledge (KA)**, **Evidence (EA)**, and a meta-cognitive **Adjudication Agent (AA)**. By first perceiving the interaction type, the framework neutralizes the implicit mandate and reasoning inertia, thereby resolving knowledge-evidence conflicts in LLM fact-checking.

3.2 Knowledge Agent (KA)

The KA functions as an autonomous component that relies exclusively on internal parametric knowledge to evaluate claim veracity. Given a claim c , the KA first selects an execution strategy $s \in \{\text{Fast}, \text{Step}\}$ via a structured prompting protocol based on the estimated reasoning complexity of c . As defined in Figure 4, Fast triggers direct inference, while Step elicits multi-step chain-of-thought reasoning. The agent then generates a binary verdict $v \in \{\text{TRUE}, \text{FALSE}\}$ and a confidence score $\delta \in [0, 1]$. To enhance calibration, we derive δ from the predictive logits of the label rather than self-reported verbalized confidence (see Appendix D for detailed analysis). Formally, the KA maps the claim c to a triplet (v, δ, s) , representing the judgment, its certainty, and the reasoning strategy.

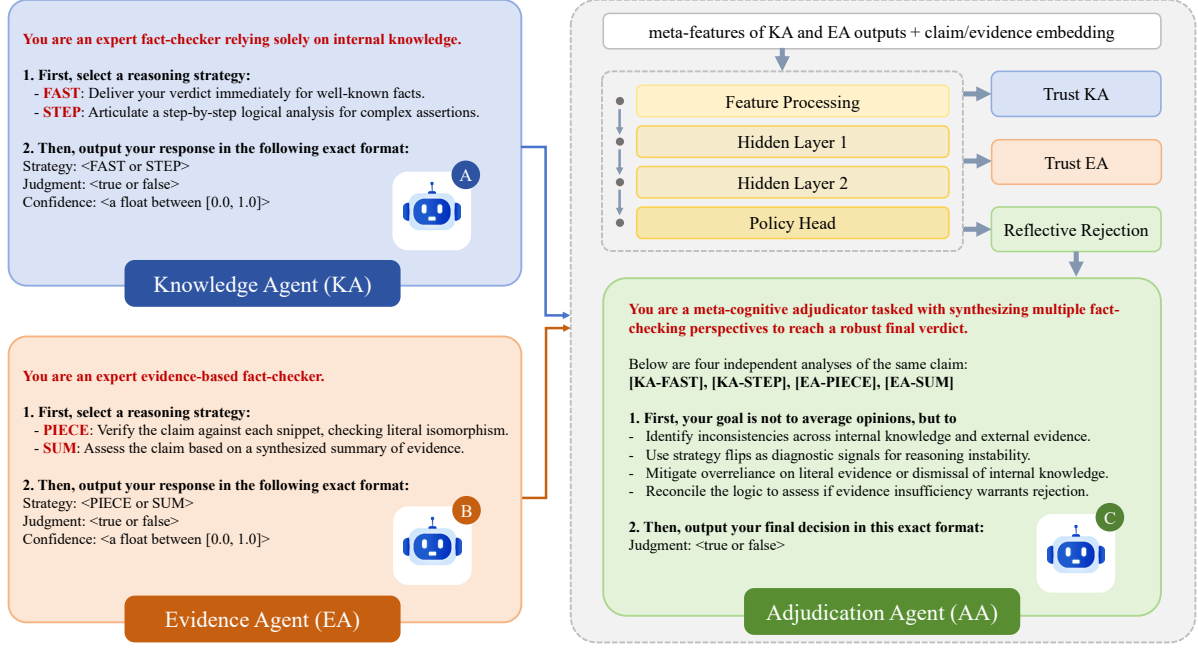


Figure 4: Framework Overview. The framework follows a “Perceive-Then-Adjudicate” policy: (i) the KA and EA provide initial judgments and confidence scores; (ii) the AA perceives the interaction type and adjudicates the agent outputs; and (iii) if necessary, the AA reconfigures the strategy to trigger Reflective Rejection for the final verdict.

3.3 Evidence Agent (EA)

The EA performs evidence-grounded verification by integrating external, human-annotated evidence from trusted sources. Given a claim c and its corresponding identifier id , the EA retrieves a pre-collected evidence set E_{id} from a curated database. These snippets are provided verbatim to the LLM to ensure high-fidelity perception of the original content. The agent then selects an execution strategy $s \in \{\text{Piece}, \text{Sum}\}$: Piece focuses on fine-grained claim-to-evidence matching, while Sum synthesizes the evidence into a holistic judgment. In a single structured response, the EA outputs a binary verdict $v \in \{\text{TRUE}, \text{FALSE}\}$ and a confidence score $\delta \in [0, 1]$. To maintain consistency with the KA, we derive δ from the predictive logits, enabling more rigorous uncertainty estimation. Formally, the EA maps the pair (c, E_{id}) to a triplet (v, δ, s) , representing the evidence-based judgment, its certainty, and the reasoning strategy employed.

3.4 Adjudication Agent (AA)

The AA acts as a meta-cognitive controller that arbitrates between the KA and EA. It operates across three distinct functional modes: TrustKA, TrustEA, and ReflectiveRejection. To determine the optimal mode, the AA constructs a holistic state vector $\mathbf{S}$ comprising (i) **meta-features** from the KA

and EA, including their respective verdicts v , confidence scores δ , strategy configurations s , and the degree of verdict consistency; and (ii) **semantic embeddings** of the claim c and evidence E_{id} . This vector $\mathbf{S}$ is processed by a trained lightweight policy network to output a discrete mode assignment.

In Reflective Rejection mode, the AA actively intervenes to break strategy rigidity by mandating bidirectional strategy switching (*e.g.*, Fast $\leftrightarrow$ Step) between alternative reasoning strategies. The task-specific agents are then compelled to regenerate their judgments using these counterpart configurations. Finally, a reflection module synthesizes the quadripartite outputs of both initial and mandated verdicts from the KA and EA to produce the definitive final verdict $v_{final} \in \{\text{TRUE}, \text{FALSE}\}$. Formally, the AA defines a mapping $\mathcal{A} : \mathbf{S} \rightarrow v_{final}$, resolving knowledge-evidence conflicts arising from implicit mandate and reasoning inertia.

3.5 Policy Learning

We optimize the AA’s policy network using a hybrid objective that integrates Proximal Policy Optimization (PPO) with Imitation Learning (IL) supervision derived from expert routing decisions. This dual-track optimization refines type-aware discernment and mitigates error propagation in high-risk scenarios characterized by joint agent unreliability.

Reward Shaping. Let y denote the ground-truth label. For a given state S , the reward $R(m)$ for selecting mode $m \in \{\text{TrustKA}, \text{TrustEA}, \text{Reject}\}$ is formulated based on the agents’ veracity and their respective confidence scores, δ_{KA} and δ_{EA} :

- **Consensus Correct:** If both agents are correct, selecting either yields $R = +\delta_m$; choosing Reject is penalized.
- **Single Agent Correct:** If only one agent is correct, selecting it yields $R = +(\delta_{correct} + (1 - \delta_{wrong}))$; other modes are penalized.
- **Consensus Wrong:** If both are incorrect, Reject yields $R = +(1 - \frac{\delta_{KA} + \delta_{EA}}{2})$, whereas selecting either agent is heavily penalized.

Safe Imitation Target. To stabilize training, we define a supervised imitation target a^{IL} : (i) $a^{IL} = \text{TrustKA}$ (or TrustEA) if only that agent is correct; (ii) $a^{IL} = \text{Reject}$ if both agents are jointly incorrect; (iii) if both are correct, a^{IL} is assigned to the agent exhibiting higher confidence.

Training Objective. The policy π_θ is optimized via a joint loss function:

$$\mathcal{L} = \mathcal{L}_{\text{PPO}} - \alpha \mathcal{H}(\pi_\theta) + \beta \mathcal{L}_{\text{CE}}(\pi_\theta, a^{IL}), \quad (1)$$

where $\mathcal{L}_{\text{PPO}}$ denotes the clipped surrogate objective, which constrains policy updates within a trust region to ensure training stability when reconciling conflicting signals. The term $\mathcal{H}(\pi_\theta)$ represents entropy regularization scaled by α to encourage exploration and prevent premature convergence to biased reasoning paths. Finally, $\mathcal{L}_{\text{CE}}$ serves as the imitation learning constraint with weight β , leveraging expert trajectories a^{IL} to provide a robust supervised signal for the adjudication logic.

4 Experiments

4.1 Datasets

We evaluate our framework on Snopes (Vlachos and Riedel, 2014) and PolitiFact (Popat et al., 2017), two publicly available benchmarks for fact-checking. Specifically, Snopes focuses on debunking contemporary rumors and urban legends, while PolitiFact primarily covers political discourse. We adopt a standard data partitioning protocol, splitting each dataset into training, validation, and test sets at a 7:1:2 ratio. Dataset statistics and representative examples are provided in Appendix A.

4.2 Experimental Setup

Baseline Models. We compare our framework with the following baselines:

- **Self-Reflect** (Shinn et al., 2023): An iterative process in which the model critiques and refines its outputs until convergence.
- **Standard Multi-Agent Debate (SMAD)** (Han et al., 2025): Two debaters with generic profiles engage in multi-turn argumentation until a judge agent makes a binary verdict.
- **Chain-of-Thought (CoT)** (Wei et al., 2022b): The KA, EA, and AA instructions are concatenated as a step-by-step reasoning process.
- **Self-Consistency** (Wang et al., 2023): This approach enhances the original CoT by aggregating multiple generated reasoning paths via majority voting to produce the final judgment.
- **Graph-based Semantic Structure with Adversarial Learning (GETRAL)** (Wu et al., 2024): This method constructs graph representations to address long-range semantic dependencies, incorporating adversarial training to enhance robustness against perturbed data.
- **Multi-view Counterfactual Contrastive Learning (MCCL)** (Zhang et al., 2024b): This method first fuses entity-level textual features with centrality-based graph features and then leverages counterfactual learning to refine the multi-view features.
- **Supervised Fine-Tuning (SFT)** (Wei et al., 2022a): This approach fine-tunes the base LLM on training data via supervised learning to specialize it for the fact-checking task.

Evaluation Metrics. We employ Accuracy (Acc) and F1-score (F1) as our metrics. To rigorously evaluate the proposed framework, we categorize the baselines into two groups. The first comprises training-free baselines (Self-Reflect, SMAD, CoT, and Self-Consistency) that operate in zero-shot settings. The second includes trained baselines (GETRAL, MCCL, and SFT), evaluated under both in-domain and out-of-domain settings. Specifically, for in-domain evaluation, models are trained and tested on the same dataset; for out-of-domain evaluation, models are trained on one dataset (e.g., Snopes) and tested on another (e.g., PolitiFact).

Method	Qwen2.5-7B-Instruct				Llama3.1-8B-Instruct			
	Snopes		PolitiFact		Snopes		PolitiFact	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Self-Reflect	0.842	0.843	0.679	0.675	0.833	0.826	0.685	0.683
SMAD	0.851	0.850	0.686	0.688	0.842	0.837	0.693	0.694
CoT	0.837	0.841	0.673	0.671	0.826	0.823	0.681	0.678
Self-Consistency	0.849	0.847	0.685	0.682	0.841	0.835	0.689	0.691
Only KA	0.730	0.658	0.486	0.351	0.724	0.698	0.506	0.432
Only EA	0.835	0.836	0.658	0.654	0.822	0.817	0.650	0.646
AA (Ours)	0.864	0.862	0.711	0.707	0.854	0.849	0.727	0.735

Table 2: Performance of training-free baselines, single agents, and our AA framework.

Model	Method	In-domain				Out-of-domain			
		Snopes		PolitiFact		Snopes		PolitiFact	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-7B-Instruct	GETRAL	0.846	0.844	0.694	0.693	0.732	0.739	0.605	0.601
	MCCL	0.853	0.855	0.706	0.706	0.782	0.786	0.639	0.632
	SFT	0.845	0.841	0.672	0.668	0.773	0.771	0.618	0.619
	AA (Ours)	0.864	0.862	0.711	0.707	0.828	0.823	0.667	0.653
Llama3.1-8B-Instruct	SFT	0.838	0.834	0.685	0.681	0.766	0.754	0.621	0.628
	AA (Ours)	0.854	0.849	0.727	0.735	0.812	0.816	0.681	0.685

Table 3: Performance of trained baselines and our AA framework in in-domain and out-of-domain settings.

Implementation Details. To ensure reproducibility, we set the temperature to 0 for all agents during the inference phase. The AA’s policy network is optimized via PPO integrated with entropy regularization to encourage diverse exploration across the action space. We provide training protocols, parameter sensitivity, and computational efficiency analyses in Appendices B, C, and E.

4.3 Main Results and Discussion

We evaluate our framework against training-free baselines (Table 2) and trained baselines under in-domain and out-of-domain settings (Table 3). Appendices F and I provide additional LLM backbones and multilingual multi-class experiments.

Performance on Training-Free Baselines. As shown in Table 2, single agents (Only KA and EA) exhibit the weakest performance, while SMAD and Self-Consistency achieve the best results among baselines. Self-Reflect consistently outperforms standard CoT, confirming the value of self-correction. Our AA framework surpasses all training-free baselines and single agents, achieving F1 improvements ranging from 1.2% to 4.1%

over the top-performing baseline. These gains highlight the advantage of meta-cognitive arbitration in resolving knowledge-evidence conflicts.

Comparison with Trained Baselines. Table 3 reports performance in both in-domain and out-of-domain settings. In-domain, specialized methods (GETRAL and MCCL) achieve strong results, and SFT performs comparably. However, these baselines suffer substantial degradation in out-of-domain: the average F1 of GETRAL drops by 9.8% and that of MCCL by 7.1%. In contrast, our AA framework maintains robust generalization with an average F1 decline of only 4.4%. Specifically, in the out-of-domain setting, our framework outperforms the strongest specialized baseline (MCCL) by an average of 3.5% in F1. This superior generalization demonstrates that our framework learns transferable arbitration strategies rather than overfitting to training distributions. Notably, the improvement is more pronounced on PolitiFact than on Snopes. This stems from a conservative bias prevalent in monolithic reasoning (see Appendix J), and our AA may be more effective at correcting such biases in domains like political fact-checking.

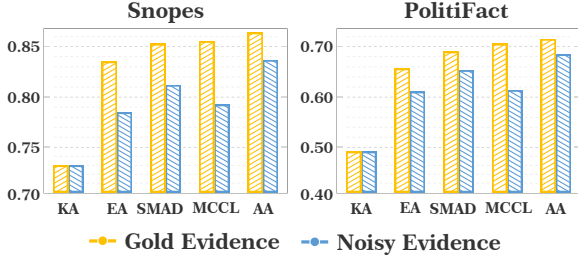


Figure 5: Accuracy of single agents, baselines, and the AA framework under noisy evidence.

4.4 Performance with Noisy Evidence

While our main experiments validate the framework’s efficacy under gold evidence conditions, practical fact-checking systems must handle data retrieved from the real world. To bridge this gap, we conduct an evaluation using the Tavily¹ search API to retrieve noisy evidence, employing Qwen2.5-7B-Instruct as the base model. For comparison, we select the top-performing models from the training-free and trained baselines discussed previously: SMAD and MCCL, respectively.

Figure 5 presents the detailed performance under noisy evidence. Compared to gold evidence, all models suffer performance degradation, confirming that evidence quality remains a challenge. Specialized methods like MCCL are vulnerable; for instance, its accuracy drops from 0.853 under gold evidence to 0.792 under noisy evidence on Snopes. In contrast, our AA framework exhibits robustness, maintaining the highest accuracy across both settings. On Snopes, AA outperforms the training-free baseline SMAD by 2.6%. On PolitiFact, this advantage over SMAD widens to 3.4%. These results confirm that the “Perceive-Then-Adjudicate” policy is effective when evidence is unreliable.

4.5 Ablation Study

We perform an ablation study across six variants to isolate the contribution of each key component:

- **Heterogeneous Agents (Heter-A):** The framework uses different LLMs (*e.g.*, KA/EA: Qwen; AA: Llama) to verify whether its effectiveness depends on agent homogeneity.
- **Confidence-based Routing (Conf-B):** The AA selects the KA or EA with higher confidence score, serving as a baseline to test if confidence alone can resolve agent conflicts.

¹<https://www.tavily.com/>

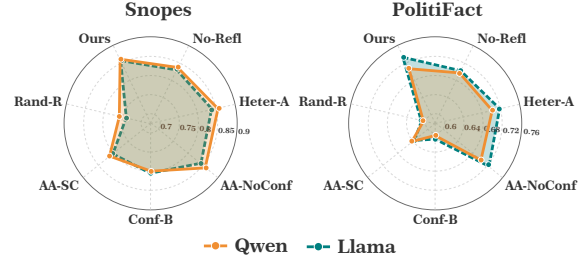


Figure 6: Ablation study of accuracy across Snopes and PolitiFact benchmarks, evaluating the full framework against six variants. “Qwen” denotes Qwen2.5-7B-Instruct; “Llama” denotes Llama3.1-8B-Instruct.

- **AA without Reflective Rejection (No-Refl):** The AA is restricted to choosing between TrustKA and TrustEA, disabling Reflective Rejection and its strategy-flipping capability.
- **Random Routing (Rand-R):** The AA selects an adjudication mode uniformly at random to evaluate the learned policy’s superiority.
- **Supervised Classifier (AA-SC):** A 3-layer MLP is trained on the same state vectors as the AA to compare hybrid learning against a supervised classifier for adjudication.
- **AA without Confidence (AA-NoConf):** Confidence scores are removed from the AA’s state vector, retaining all other features to assess the degree of reliance on confidence.

The full framework consistently outperforms all variants across both datasets in Figure 6. The **Heter-A** variant achieves the highest performance among ablations, demonstrating that our framework learns a model-agnostic arbitration policy capable of independent judgment regardless of agent homogeneity. The minimal performance drop of **AA-NoConf** indicates that the AA’s decisions do not exclusively rely on raw confidence scores and are robust to potential miscalibration in LLM logits. Conversely, the poor performance of **Conf-B** reveals that simple confidence heuristics are insufficient to resolve complex agent conflicts. Furthermore, **AA-SC** underperforms the full framework, showing that hybrid learning discovers more effective arbitration strategies than a supervised classifier. The inferior results of **Rand-R** validate that the learned policy captures task-specific patterns rather than relying on chance. Finally, the competitive yet lower performance of **No-Refl** confirms that strategy-flipping is vital for error correction.

Types	Snopes			PolitiFact		
	Hit	Total	%	Hit	Total	%
<i>B-C</i>	547	554	98.7	267	277	96.4
<i>K-R</i>	47	80	58.8	40	70	57.1
<i>E-R</i>	143	165	86.7	165	193	85.5
<i>B-W</i>	19	70	27.1	44	174	25.3

Table 4: Adjudication performance across four types. For each type, **Total** indicates the sample count, **Hit** denotes the cases correctly adjudicated by the AA, and % represents the percentage of correctly handled cases.

4.6 Per-Type Accuracy Analysis

To validate the framework’s effectiveness, we analyze its performance across four distinct agent interaction types using Qwen2.5-7B-Instruct as the backbone. Table 4 shows that the AA maintains high fidelity on **Both-Correct (B-C)** cases, achieving an average accuracy of 97.6%. This confirms that the framework effectively preserves reliable consensus without unnecessary disruption. In instances of conflict, the AA avoids over-reliance on external evidence through sophisticated adjudication logic. While the EA generally exhibits higher baseline accuracy, our framework transcends a simplistic evidence-first heuristic by achieving over 57% accuracy on **KA-Reliable (K-R)** cases and an average of 86.1% accuracy on **EA-Reliable (E-R)** cases. This calibrated balance reflects the AA’s ability to counteract the tendency to misinterpret neutral instructions as an implicit mandate to prioritize evidence at all costs. The advantage of the type-aware policy is most salient in **Both-Wrong (B-W)** cases, where the AA leverages Reflective Rejection to trigger strategy flipping. This intervention recovers correct judgments in 26.2% of joint-failure cases on average, salvaging samples that would otherwise succumb to an erroneous consensus. An error case study is in Appendix L.

5 Related Work

Automated fact-checking has traditionally centered on aligning claims with supporting or refuting external evidence. Early neural approaches operationalized this paradigm through learned representations; for instance, Popat et al. (2018) combined linguistic features with source credibility, while Ma et al. (2019) employed hierarchical attention to learn claim-guided evidence representations. However, these methods often frame fact-checking as a classification task over fixed evidence sets, and were thus restricted by shallow semantic modeling.

Subsequent research introduced richer structural interactions, notably through graph-based frameworks that capture long-range semantic dependencies via neighborhood propagation (Xu et al., 2022). Systems like Wu et al. (2024) further enhanced representation robustness through supervised contrastive learning and adversarial augmentation.

The field has recently pivoted toward LLMs as versatile reasoning agents. Lin et al. (2025) proposed an adaptive multi-agent system that iteratively evaluates challenging cases, while Mujahid et al. (2025) demonstrated that aligning prompts with expert guidelines improves reliability in politically sensitive contexts. Furthermore, Devasier et al. (2025) integrated LLMs with structured databases using frame semantics for precise retrieval. Despite such sophistication, existing systems typically apply fixed reasoning strategies and aggregate judgments without actively managing the inherent cognitive conflict between internal knowledge and external evidence. Consequently, they remain prone to high-confidence erroneous consensus and an asymmetric reliance on grounded evidence. In contrast, our framework transcends simple evidence incorporation by introducing dynamic multi-agent arbitration to resolve these conflicts and rectify systemic reasoning failures.

6 Conclusion

This paper introduces a **Multi-Agent Arbitration Framework** for automated fact-checking, designed to resolve cognitive conflicts and erroneous consensus between internal knowledge and external evidence. Unlike monolithic or static fusion paradigms, our framework employs an AA, optimized through reinforcement and imitation learning, to dynamically mediate between a KA and an EA. Experimental results on Snopes and PolitiFact benchmarks demonstrate that the AA consistently outperforms baselines, including SFT, CoT, and specialized methods. Specifically, our **“Perceive-Then-Adjudicate”** policy mitigates the bias induced by **implicit mandate**, while Reflective Rejection enables the framework to rectify 26.2% of challenging erroneous consensus cases stemming from **reasoning inertia**. These results indicate that our design endows the system with inherent self-calibration capabilities, ensuring the robust integration of external information without sacrificing internal judgmental autonomy, thereby establishing a foundation for reliable fact-checking frameworks.

Limitations

Despite the observed performance gains, our framework presents several limitations. First, our scope is limited to textual data, whereas modern disinformation increasingly involves manipulated images, videos, or audio. Extending the multi-agent arbitration framework to process multimodal data is a promising direction for enhancing the framework’s versatility. Second, although the AA enables error recovery, the increased computational overhead incurred by Reflective Rejection warrants further optimization for real-time deployment. In summary, our analysis of knowledge–evidence conflicts lays the groundwork for developing more robust, self-correcting automated fact-checking frameworks.

Ethical Considerations

This work enhances the reliability and cognitive autonomy of LLMs in automated fact-checking, directly addressing the societal challenge of misinformation. Our framework is designed to mitigate compliance-induced errors and does not introduce new risks related to the generation of harmful or deceptive content. The main experiments use publicly available, human-annotated benchmarks consisting of curated claims and evidence that contain no personally identifiable information or sensitive user data. For the experiments involving noisy evidence, the real-world data retrieved via the Tavily Search API was processed in strict accordance with the provider’s Terms of Service and for its intended purpose, ensuring that only publicly accessible web content was used for academic evaluation without violating individual privacy or infringing copyright. While the framework assumes access to authoritative evidence, we acknowledge that real-world sources may inherently reflect societal biases. Therefore, any future deployment must involve rigorous auditing of both claim and evidence sources to ensure fairness, representativeness, and transparency. This research contributes to the development of robust and self-aware AI systems for trustworthy truth assessment. During the preparation of this manuscript, the authors used the Qwen model to assist with language polishing and grammar refinement to improve readability. All intellectual contributions, experimental designs, and interpretations of the results were developed and verified by the authors. The authors assume full responsibility for the content of the final manuscript.

Acknowledgments

The work of the Wuhan University authors was supported by Ant Group through the CCF-Ant Research Fund and by the National Natural Science Foundation of China (No. 62176187). Zhuang Li received no funding for this work. Fei Li and Donghong Ji are co-corresponding authors. Zhuang Li is the senior author.

References

- Marah I Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat S. Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. [Phi-4-reasoning technical report](#). *CoRR*, abs/2504.21318.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, and 3 others. 2021. [A general language assistant as a laboratory for alignment](#). *CoRR*, abs/2112.00861.
- Yingjian Chen, Haoran Liu, Yinlong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, and Irene Li. 2025. [Graphcheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14976–14995. Association for Computational Linguistics.
- Jacob Daniel Devasier, Akshith Reddy Putta, Rishabh Mediratta, and Chengkai Li. 2025. [Task-oriented automatic fact-checking with frame-semantics](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13825–13842. Association for Computational Linguistics.
- Jianglin Dong, Haixia Mao, Yiyi Zhao, Yuan Peng, Junyi Yang, and Jiangping Hu. 2024. [A cognitive inertia sequence model for opinion formation in group decision-making systems](#). In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2024, Kuching, Malaysia, October 6-10, 2024*, pages 348–353. IEEE.
- Fabio Egle, Felix Kluge, Daniel Schoene, Linda Becker, and Anne D. Koelewijn. 2022. [Development of an inertial sensor-based exergame for combined cognitive and physical training](#). In *IEEE-EMBS International Conference on Wearable and Implantable*

- Body Sensor Networks, BSN 2022, Ioannina, Greece, September 27-30, 2022*, pages 1–4. IEEE.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. [Large language model based multi-agents: A survey of progress and challenges](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8048–8057. ijcai.org.
- Ashim Gupta and Vivek Srikumar. 2021. [X-factor: A new benchmark dataset for multilingual fact checking](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 675–682. Association for Computational Linguistics.
- Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. [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, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 15114–15129. Association for Computational Linguistics.
- Xinyi Jiang, Tianyi Hu, Yuheng Qin, Guoming Wang, Zhou Huan, Kehan Chen, Gang Huang, Rongxing Lu, and Siliang Tang. 2025. [Chatmap: Mining human thought processes for customer service chatbots via multi-agent collaboration](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11927–11947. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislaw Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *CoRR*, abs/2207.05221.
- Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. [Debating with more persuasive llms leads to more truthful answers](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. [Confidence under the hood: An investigation into the confidence-probability alignment in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 315–334. Association for Computational Linguistics.
- Choonsik Lee and Matthew E. Souther. 2020. [Managerial reliance on the retail shareholder vote: Evidence from proxy delivery methods](#). *Manag. Sci.*, 66(4):1717–1736.
- 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 the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 360–381. Association for Computational Linguistics.
- Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Zhiwei Steven Wu, Bo Li, and Ding Zhao. 2022. [Constrained variational policy optimization for safe reinforcement learning](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of Proceedings of Machine Learning Research*, pages 13644–13668. PMLR.
- Jing Ma, Wei Gao, Shafiq R. Joty, and Kam-Fai Wong. 2019. [Sentence-level evidence embedding for claim verification with hierarchical attention networks](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2561–2571. Association for Computational Linguistics.
- 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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. [Factual confidence of llms: on reliability and robustness of current estimators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4554–4570. Association for Computational Linguistics.
- Zain Muhammad Mujahid, Dilshod Azizov, Maha Tufail Agro, and Preslav Nakov. 2025. [Profiling news media for factuality and bias using llms and the fact-checking methodology of human experts](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*.

- pages 798–819. Association for Computational Linguistics.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. [Discovering language model behaviors with model-written evaluations](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics.
- Kashyap Papat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. [Where the truth lies: Explaining the credibility of emerging claims on the web and social media](#). In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 1003–1012. ACM.
- Kashyap Papat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. [Declare: Debunking fake news and false claims using evidence-aware deep learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 22–32. Association for Computational Linguistics.
- Morgane Rivi re, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L onard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram , Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *CoRR*, abs/2408.00118.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Andreas Vlachos and Sebastian Riedel. 2014. [Fact checking: Task definition and dataset construction](#). In *Proceedings of the Workshop on Language Technologies and Computational Social Science@ACL 2014, Baltimore, MD, USA, June 26, 2014*, pages 18–22. Association for Computational Linguistics.
- Minzheng Wang, Xinghua Zhang, Kun Chen, Nan Xu, Haiyang Yu, Fei Huang, Wenji Mao, and Yongbin Li. 2025. [DEMO: reframing dialogue interaction with fine-grained element modeling](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11373–11401. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. [Finetuned language models are zero-shot learners](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Junfei Wu, Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2024. [Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks](#). *IEEE Trans. Knowl. Data Eng.*, 36(11):5591–5604.
- Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. [Evidence-aware fake news detection with graph neural networks](#). In *WWW ’22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2501–2510. ACM.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang,

Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. [Chatglm: A family of large language models from GLM-130B to GLM-4 all tools](#). *CoRR*, abs/2406.12793.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. [LUQ: long-text uncertainty quantification for llms](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5244–5262. Association for Computational Linguistics.

Yongcheng Zhang, Lingou Kong, Sheng Tian, Hao Fei, Changpeng Xiang, Huan Wang, and Xiaomei Wei. 2024b. [Multi-view counterfactual contrastive learning for fact-checking fake news detection](#). In *Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, June 10-14, 2024*, pages 385–393. ACM.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2025. [Conformity in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3854–3872. Association for Computational Linguistics.

A Dataset Details

A.1 Data Statistics and Partitioning

We evaluate our framework using two prominent fact-checking benchmarks: Snopes² and PolitiFact³. Snopes is one of the longest-running and most comprehensive fact-checking websites, primarily focusing on urban legends, internet rumors, and cultural myths that circulate widely on social media. In contrast, PolitiFact specializes in political discourse, verifying claims made by elected officials, candidates, and political parties. To ensure a rigorous evaluation, we follow a 7:1:2 split for training, validation, and testing. As shown in Table 5, the datasets contain a substantial amount of external evidence. Notably, Snopes exhibits a higher degree of class imbalance, with 2.7 times as many FALSE claims as TRUE claims, whereas PolitiFact presents a more balanced distribution.

A.2 Data Format and Field Definitions

We adopt the relational data structure defined in Table 6 to standardize the inputs from these benchmarks. Each instance is organized as a one-to-many mapping, in which `id_left` serves as the unique claim identifier and `id_right` indexes individual evidence snippets to maintain structural alignment.

²<https://www.snopes.com/>

³<https://www.politifact.com/>

Dataset	# TRUE	# FALSE	# Total	# Evidence
Snopes	1,164	3,177	4,341	29,242
PolitiFact	1,867	1,701	3,568	29,556

Table 5: Statistics of the experimental datasets. The symbol “#” denotes the count of claims (categorized by ground-truth labels) and the number of evidence snippets across training, validation, and test sets.

The `claim_text` and `evidence` fields provide the raw textual content for semantic encoding, while `cred_label` serves as the supervision signal for veracity prediction. Additionally, the dataset includes the `claim_source` and `evidence_source` fields, which jointly and explicitly document the speaker’s identity and the publisher’s domain for each claim and evidence snippet, respectively.

B Implementation Details

B.1 State Representation and Composition

The state for the Adjudication Agent (AA) is synthesized from the intermediate outputs of the Knowledge Agent (KA) and the Evidence Agent (EA). This multifaceted representation serves as the informational foundation for the AA to execute meta-reasoning and resolve cross-agent conflicts.

- **Semantic Encoding:** We extract semantic representations from the last-layer final-token position. The feature dimension is intrinsic to each architecture: 4,096 for Llama3.1-8B-Instruct and 3,584 for Qwen2.5-7B-Instruct.
- **Execution Metadata:** To enrich semantic representations with decision context, we concatenate the hidden states with a set of metadata features derived from the KA and EA:
 1. **Confidence Scores:** Scalar values in $[0, 1]$ derived from output token logits as specified in Section 3, quantifying the agents’ internal epistemic certainty.
 2. **Prediction Alignment:** A binary indicator encoding the consensus between the KA and EA regarding the label.
 3. **Predicted Labels:** The independent veracity labels predicted by KA and EA.
 4. **Strategy Selection:** Categorical encodings that encapsulate the distinct reasoning pathways selected in prior stages, such as the KA’s FAST vs. STEP modes or the EA’s SUM vs. PIECE strategies.

Field	Example Value	Description
<i>Claim-level Fields</i>		
id_left	0	Unique identifier for the claim-level instance.
cred_label	TRUE	Ground-truth veracity label for the claim.
claim_text	<i>federal tax code loopholes...</i>	The verbalized statement to be verified.
claim_source	<i>Eason</i>	The originator or speaker of the claim.
<i>Evidence-level Fields</i>		
id_right	0	Index for an individual evidence snippet associated with a claim.
evidence	<i>for firms moving overseas...</i>	Supporting text retrieved from external sources.
evidence_source	<i>senate.gov</i>	The original publisher or domain of the evidence.

Table 6: Data schema and field definitions. The structure facilitates a relational mapping between a unique claim and its retrieved evidence pool.

B.2 Training Protocols and Hardware

We instantiate our Multi-Agent Arbitration Framework using **Qwen2.5-7B-Instruct** and **Llama3.1-8B-Instruct** as the backbones.

- **Network Architecture:** The policy network follows a four-layer design: (i) a linear projection from the input dimension to 256; (ii) a hidden layer from 256 to 128 with ReLU and Dropout; (iii) a hidden layer from 128 to 128 with ReLU and Dropout; and (iv) a policy head from 128 to 3 logits for TrustKA, TrustEA, and Reject, followed by Softmax.
- **Hyperparameter Settings:** For all reinforcement learning phases, we maintain a total batch size of 32. The initial learning rate is set to 1.0×10^{-5} , governed by a cosine annealing schedule with a 0.1 warm-up ratio to stabilize policy gradient updates and mitigate premature convergence to local optima.
- **Validation and Model Selection:** To mitigate overfitting and ensure generalizability, we sequester 10% of the entire dataset as an independent validation set, which is strictly excluded from weight optimization. Model selection is performed based on the optimal validation loss and task-specific metrics (e.g., Accuracy) observed across 10 training epochs.
- **Hardware Infrastructure:** All experiments were conducted on a dedicated high-performance compute node equipped with a 24-core CPU, 240 GB of RAM, and two NVIDIA A100 (80GB) GPUs. This infrastructure accommodates the substantial memory footprint of experience buffers during PPO updates and supports efficient large-scale model inference. We implemented all experiments using PyTorch 2.1.0 and Transformers 4.36.0.

- **Statistical Robustness:** To ensure the reliability of our findings, all reported results are averaged over 3 independent runs with distinct random seeds. We report the mean performance across these trials to account for potential variance in the hybrid learning phases.

C Parameter Sensitivity Analysis

We evaluate the sensitivity of core hyperparameters governing our hybrid objective function:

$$\mathcal{L} = \mathcal{L}_{\text{PPO}} - \alpha \mathcal{H}(\pi_{\theta}) + \beta \mathcal{L}_{\text{CE}}\{\pi_{\theta}, a^{IL}\}, \quad (2)$$

where $\mathcal{L}_{\text{PPO}}$ is the clipped surrogate objective, $\mathcal{H}$ denotes entropy regularization, and $\mathcal{L}_{\text{CE}}$ represents a class-weighted cross-entropy loss. The coefficients α and β control the strength of policy exploration and supervised imitation, respectively.

PPO Clipping Parameter (ϵ) The clipping parameter ϵ serves as a safeguard against excessively large policy updates. Our sensitivity analysis, summarized in Figure 7, reveals that the classification accuracy for both Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct peaks at $\epsilon = 0.2$. At this threshold, policy updates are sufficiently large to dynamically adapt to evidence while remaining adequately constrained to prevent the potential catastrophic forgetting of learned reasoning patterns.

Entropy Regularization Coefficient (α) The entropy coefficient α is pivotal for modulating the exploration-exploitation trade-off. As illustrated in Figure 8, our empirical analysis on Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct across both benchmarks demonstrates that α dictates the AA’s exploration propensity toward the Reflective Rejection pathway. When α is negligible (< 0.001), the policy tends to collapse into a sub-optimal greedy

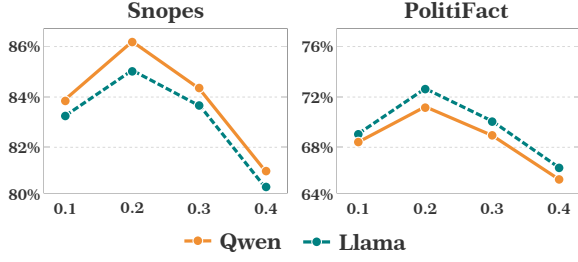


Figure 7: Performance variation of Qwen2.5-7B-Instruct (Qwen) and Llama3.1-8B-Instruct (Llama) across PPO clipping parameters ϵ . Accuracy peaks at $\epsilon = 0.2$, striking an optimal balance between policy plasticity and the preservation of reasoning consistency.

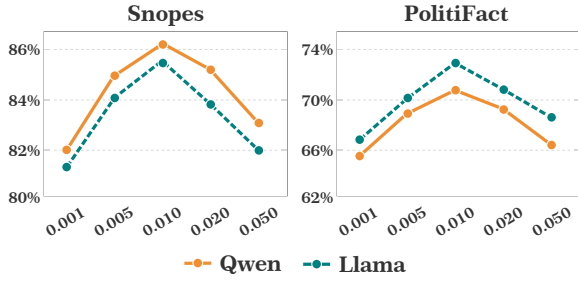


Figure 8: Impact of Entropy Coefficient α on accuracy for Qwen2.5-7B-Instruct (Qwen) and Llama3.1-8B-Instruct (Llama). A trend is observed, indicating that moderate exploration ($\alpha \approx 0.01$) is essential for the AA to invoke the Reflective Rejection mechanism.

strategy. Conversely, maintaining $\alpha = 0.01$ preserves sufficient policy entropy, facilitating the exploration of intensive reflection mechanisms.

Imitation Learning Weight (β) The imitation weight β anchors reinforcement learning to a supervised baseline a^{LL} . As depicted in Figure 9, while $\beta = 0$ (pure RL) leads to suboptimal accuracy due to sparse reward signals, a setting of $\beta = 0.5$ yields a synergistic equilibrium between supervised stability and reinforcement-driven refinement. Excessively high β values, however, stifle agent autonomy, causing the models to merely mimic the training distribution, which results in a performance plateau or decline in overall accuracy.

D Confidence Calibration Analysis

A potential concern regarding our framework is that the AA’s arbitration relies on confidence scores, which are poorly calibrated in instruction-tuned LLMs (Kumar et al., 2024; Mahaut et al., 2024). To validate our design, we conduct a calibration analysis comparing logit-based confidence (Logit) with verbalized self-reported confidence (Self).

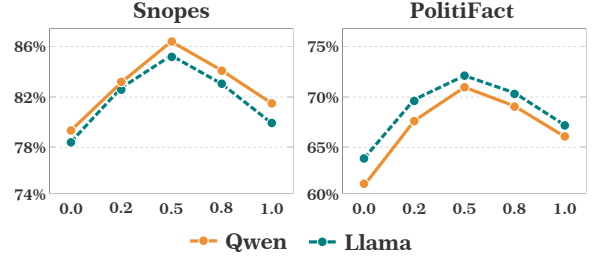


Figure 9: Sensitivity of accuracy to Imitation Weight β for Qwen2.5-7B-Instruct (Qwen) and Llama3.1-8B-Instruct (Llama). The configuration ($\beta = 0.5$) yields peak performance, mitigating the training instability inherent in reinforcement learning during optimization.

Table 7 presents the Expected Calibration Error (ECE) for both the KA and EA. A lower ECE indicates a closer alignment between confidence estimates and actual accuracy. As shown, Logit consistently achieves substantially lower ECE than Self across all models and datasets. For instance, on Snopes using Qwen2.5-7B-Instruct, the EA’s ECE drops from 0.245 (Self) to 0.082 (Logit), demonstrating that logit-based estimates provide a relatively reliable foundation for adjudication. However, given that Logit exhibits moderate miscalibration, the question remains whether the AA’s routing decisions are stable. Our studies in Section 4.5 address this concern from two perspectives:

- **Insufficiency of Heuristics:** The poor performance of **Conf-B** (Confidence-based Routing) reveals that simple heuristics relying solely on confidence scores fail to resolve complex agent conflicts. This underscores that raw confidence, regardless of its calibration, is insufficient for reliable fact-checking arbitration.
- **Holistic State Representation:** The AA’s state vector $\mathbf{S}$ comprises not only confidence scores but also agent verdicts, strategy selections, prediction consistency, and semantic embeddings. This holistic representation enables the AA to perceive relational patterns rather than merely react to absolute confidence values. Notably, the **AA-NoConf** (AA without Confidence) variant, which removes confidence scores entirely, suffers a minimal performance drop. This demonstrates that the AA’s arbitration capability does not solely depend on precise calibration, as the framework remains effective even without this signal.

Model	Dataset	Confidence Type	EA-ECE	KA-ECE
Qwen2.5-7B-Instruct	Snopes	Self	0.245	0.312
		Logit	0.082	0.115
	PolitiFact	Self	0.268	0.334
		Logit	0.094	0.128
Llama3.1-8B-Instruct	Snopes	Self	0.213	0.289
		Logit	0.076	0.104
	PolitiFact	Self	0.235	0.306
		Logit	0.088	0.118

Table 7: Expected Calibration Error (ECE) of logit-based (Logit) and verbalized self-reported (Self) confidence for KA and EA. Lower ECE indicates better calibration.

Dataset	Method	Latency (ms/sample)	Acc (%)
Snopes	KA	190	73.0
	EA	290	83.5
	CoT	320	83.7
	Ours	400	86.4
PolitiFact	KA	200	48.6
	EA	300	65.8
	CoT	330	67.3
	Ours	420	71.1

Table 8: Computational overhead comparison on Snopes and PolitiFact using Qwen2.5-7B-Instruct. Latency is measured in milliseconds per sample with model parallelism.

E Computational Efficiency Analysis

A concern for our fact-checking framework is the computational cost introduced by multi-agent interaction and potential reruns from Reflective Rejection. To address this, we quantify the inference latency and computational overhead using Qwen2.5-7B-Instruct as the backbone. Latency represents the end-to-end processing time per sample.

Table 8 summarizes the overhead. Our framework incurs $1.25\text{--}1.27\times$ the latency of CoT, primarily due to the parallel execution of the KA and EA. This investment yields a favorable accuracy-efficiency trade-off: on Snopes, 25% higher end-to-end latency delivers a 2.7% accuracy gain (approximately 9.3% overhead per 1% improvement). On the PolitiFact benchmark, 27% higher end-to-end latency achieves a 3.8% gain (7.1% overhead per 1% improvement). This trade-off is justifiable given the societal stakes of misinformation. Moreover, Reflective Rejection triggers in only 15.4% of cases on average, concentrating deeper reasoning on ambiguous samples where it provides the

most value, while the remaining cases incur no additional reflection-specific rerun cost.

We analyze memory usage and throughput. Our framework has a peak memory usage of around 24.8 GB, compared to roughly 20 GB for CoT, and achieves a throughput of 2.38-2.50 samples/sec, compared to 3.03-3.13 samples/sec for CoT. This overhead is justified by the 2.7-3.8% accuracy gain. For deployment, one can reduce the number of reflection steps or use smaller backbones.

F Additional LLM Backbones Analysis

In our main experiments, we focus on Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct as backbone models for the evaluation of our framework and its variants. To further demonstrate the generalizability of our framework, we conduct additional experiments on a wider range of LLM backbones, including ChatGLM2-6B (Zeng et al., 2024), Gemma-3-12B (Rivière et al., 2024), Qwen2.5-72B (Yang et al., 2024), Llama3.1-70B-Instruct (Team, 2024), Phi-4-Mini-Reasoning (Abdin et al., 2025), and GLM-Z1-32B-0414 (Zeng et al., 2024).

Category	Model	Method	Snopes		PolitiFact		
			Acc	F1	Acc	F1	
Open-source LLMs	ChatGLM2-6B	KA	0.731	0.618	0.479	0.312	
		EA	0.735	0.627	0.487	0.332	
	Gemma-3-12B	KA	0.732	0.619	0.478	0.309	
		EA	0.736	0.621	0.479	0.312	
	Qwen2.5-72B	KA	0.730	0.688	0.513	0.428	
		EA	0.855	0.859	0.681	0.681	
	Llama3.1-70B-Instruct	KA	0.747	0.703	0.539	0.473	
		EA	0.814	0.823	0.699	0.695	
	Reasoning Models	Phi-4-Mini-Reasoning	KA	0.727	0.628	0.486	0.332
			EA	0.776	0.726	0.546	0.462
GLM-Z1-32B-0414		CoT	0.795	0.742	0.602	0.528	
		KA	0.731	0.624	0.479	0.321	
		EA	0.733	0.625	0.481	0.326	
		CoT	0.752	0.649	0.525	0.401	

Table 9: Extended results on diverse LLM backbones. Performance of KA and EA is reported across various open-source LLMs and reasoning models on Snopes and PolitiFact. For reasoning models, CoT results are provided.

Table 9 presents the performance of KA and EA across these models. For reasoning-specialized models (Phi-4-Mini-Reasoning and GLM-Z1-32B-0414), we also include CoT results. These extended results further corroborate our main findings: (i) evidence augmentation consistently improves performance across diverse LLM backbones; (ii) model scale contributes to accuracy but does not eliminate the need for structured arbitration; and (iii) our AA framework, even with smaller backbones, achieves competitive or superior performance compared to larger models with standard prompting.

G Extended Ablation Study

In order to verify the effectiveness of hybrid learning, we conduct ablation experiments on Snopes and PolitiFact using Qwen2.5-7B-Instruct (Qwen) and Llama3.1-8B-Instruct (Llama) as backbones, with accuracy as the metric. The two variants are as follows: the PPO-only (P-O) variant removes the imitation learning term, relying solely on reinforcement learning with entropy regularization; the IL-only (I-O) variant removes the PPO term and optimizes the routing policy using only supervised learning. The results show that our hybrid learning objective outperforms both variants across all backbones and datasets, confirming that combining reinforcement and imitation learning is beneficial.

Model	Snopes		PolitiFact	
	Qwen	Llama	Qwen	Llama
P-O	0.834	0.828	0.665	0.679
I-O	0.810	0.796	0.650	0.657
Ours	0.864	0.854	0.711	0.727

Table 10: Extended ablation study results (Accuracy) of hybrid learning across different backbones and datasets.

H Extended Per-type Analysis

Table 11 reports per-type accuracy (%) on Snopes for SMAD, MCCL, our framework, and its ablations. Our full framework and its ablations use Qwen2.5-7B-Instruct. We categorize instances into four types: *Both-Correct* (B-C), *KA-Reliable* (K-R), *EA-Reliable* (E-R), and *Both-Wrong* (B-W). Our framework outperforms SMAD and MCCL, confirming that combining reinforcement and imitation learning helps the AA identify conflict signals for fact-checking arbitration. None of the baselines outperform our framework, demonstrating the effectiveness of our Reflective Rejection mechanism on B-W instances. Among the ablations, the variant without Reflective Rejection (No-Refl) drops sharply on B-W, confirming that Reflective Rejection is important for breaking erroneous consensus. Random Routing (Rand-R) performs poorly across all types, validating that the learned policy captures meaningful patterns rather than relying on chance.

Models	B-C	K-R	E-R	B-W
SMAD	98.3	51.2	85.9	17.1
MCCL	98.5	50.0	85.4	20.0
Heter-A	98.5	52.5	86.0	21.4
Conf-B	97.1	28.7	80.6	5.7
No-Refl	98.1	46.2	84.8	14.2
Rand-R	95.1	20.0	68.4	2.8
AA-SC	97.4	35.0	81.8	8.5
AA-NoConf	98.3	53.7	83.6	24.2
Ours	98.7	58.8	86.7	27.1

Table 11: Extended per-type accuracy (%) analysis on Snopes. Models include the top-performing baselines from Section 4.3 and the ablations of our framework from Section 4.5. B-C, K-R, E-R, and B-W denote *Both-Correct*, *KA-Reliable*, *EA-Reliable*, and *Both-Wrong*.

Additionally, we provide an error breakdown of misclassified B-W cases, comprising 51 Snopes cases and 130 PolitiFact cases. On PolitiFact, the majority of errors (107 cases, 82.31%) are false negatives (where both the KA and EA predict FALSE while the ground-truth label is TRUE), revealing a conservative bias, as discussed in Appendix J. In contrast, the distribution on Snopes is more balanced, with 34 false negatives (66.67%) and 17 false positives (33.33%), likely due to its broader topical scope. To complement this failure analysis and evaluate potential false triggering or mistaken disruption, we analyze B-C instances (554 for Snopes and 277 for PolitiFact). Reflective Rejection is triggered on only 8.4% of Snopes B-C instances and 9.3% of PolitiFact B-C instances. Crucially, the vast majority of these triggered cases still yield a correct final verdict, with actual mistaken disruption occurring in only 1.3% of Snopes cases and 3.6% of PolitiFact cases. This confirms that correct consensus is rarely disrupted, demonstrating that our conflict perception and adjudication framework is both conservative and effective.

I Cross-Lingual and Multi-Class

To assess our framework in multilingual and multi-class fact-checking, we conduct additional experiments on the X-Fact benchmark (Gupta and Sriku-mar, 2021), which includes 7-class veracity labels across diverse languages. Specifically, the KA and EA each output a 7-way verdict rather than binary TRUE/FALSE, while the AA performs its original arbitration task: choosing TrustKA, TrustEA, or Reject. Since the AA’s role is to arbitrate between

agents rather than to classify, its binary-trained policy generalizes to the multi-class setting.

Table 12 summarizes the results, from which several key observations emerge. First, the AA consistently outperforms both the KA and EA across all languages and backbones. For instance, on the German subset with Qwen2.5-7B-Instruct, the AA achieves a 5.6% accuracy gain over the stronger single agent, which demonstrates that our arbitration mechanism remains effective beyond English binary classification. Second, the relative performance ordering mirrors our main findings on Snopes and PolitiFact, indicating that the “Perceive-Then-Adjudicate” policy captures generalizable arbitration principles that transfer across languages. Finally, both Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct exhibit similar patterns across languages, suggesting that our framework relies on fundamental reasoning about knowledge-evidence conflicts rather than language-specific features.

J Conservative Bias Analysis

To investigate the systematic skew toward false-negative (FN) predictions, which we conceptualize as conservative bias in Section 4.3, we perform a granular error decomposition on the subset of misclassified instances. As shown in Table 13, the KA configurations reveal a profound skepticism baseline. In the absence of external grounding, the vast majority of model errors are concentrated in the FN category, meaning truthful claims are incorrectly rejected as FALSE. This tendency is most pronounced on PolitiFact, where FN errors account for 97.2% of total misclassifications for Qwen2.5-7B-Instruct and 91.9% for Llama3.1-8B-Instruct. This suggests that parametric knowledge across different architectures is often insufficient for verification, leading models to prioritize caution and default to a negative verdict when their training memory lacks explicit cross-references.

The integration of the EA recalibrates this error distribution, though the magnitude varies by domain and architecture. For Qwen2.5-7B-Instruct on Snopes, the share of false negatives among all errors decreases by 40.3 percentage points, from 87.17% for the KA to 46.92% for the EA. However, a substantial conservative bias persists on PolitiFact, where over 68.4% of errors remain categorized as FN despite external evidence availability. This disparity suggests the bottleneck in political fact-checking is not merely an informational deficit,

Model	Method	German		Portuguese	
		Acc	F1	Acc	F1
Qwen2.5-7B-Instruct	KA	0.282	0.215	0.254	0.186
	EA	0.385	0.342	0.362	0.315
	AA	0.441	0.412	0.415	0.382
Llama3.1-8B-Instruct	KA	0.304	0.238	0.276	0.202
	EA	0.372	0.325	0.355	0.301
	AA	0.432	0.398	0.428	0.389

Table 12: Performance comparison on the X-Fact benchmark across German and Portuguese (7-class veracity labels). The AA consistently outperforms both the KA and the EA across languages and backbone models.

but one of evidentiary stringency. Models maintain an excessively high verification threshold for political discourse, opting for safe but incorrect rejections when the mapping between relevant retrieved facts and claims lacks literal isomorphism.

K Detailed Prompt Templates

To ensure reproducibility and facilitate fair comparison across different prompting paradigms, we provide the standardized templates used to instantiate our modular components. While our primary framework operates by calling these modules independently via an autonomous agent, we also aggregate these stages into a unified Chain-of-Thought (CoT) sequence for baseline evaluations. This sequential formatting demonstrates how reasoning, retrieval, and meta-level classification are orchestrated within a single prompt trajectory.

As detailed in Table 14, the templates define the specific functional roles of each stage: (1) the **Knowledge Agent** leverages internal parametric memory; (2) the **Evidence Agent** grounds its judgments in external evidence; (3) the **Conflict Perception** module serves as a meta-reasoning gatekeeper for categorizing agent interactions, although its prompt is provided only as an interpretive template, with actual routing decisions produced by the learned policy network; and (4) the **Reflective Rejection** module synthesizes counterfactual perspectives. By enforcing structured outputs and explicit strategy selection (*e.g.*, FAST vs. STEP), this CoT formulation highlights the disentanglement between internal beliefs and external evidence, even when the stages are executed sequentially.

L Error Case Study

To investigate the limitations of our framework, we conduct an error analysis by examining three

representative failure cases across distinct types: *KA-Reliable*, *EA-Reliable*, and *Both-Wrong*. Here, *[Summary]* indicates that we sample error cases and instruct each agent (KA, EA, and AA) to provide a concise explanation alongside its verdict. These explanations are manually verified and lightly edited to pinpoint the causes of erroneous decisions while ensuring clarity and faithfulness.

L.1 KA-Reliable

Table 15 presents a *KA-Reliable* failure case involving the claim that “The tech company banned all remote work globally in early 2018.” The KA correctly identifies this claim as FALSE based on its internal parametric knowledge of the company’s adoption and expansion of hybrid work arrangements. However, the EA is misled by sensational social media snippets that misinterpret a leaked internal memo concerning localized restrictions as an actual global ban, thereby leading to an incorrect TRUE verdict. The AA fails to resolve this conflict properly; instead of validating the factual knowledge of the KA, it over-relies on the external text and selects TrustEA. It wrongly assumes that emotionally charged social media buzz outweighs parametric knowledge, highlighting the framework’s inherent vulnerability to noisy external evidence.

L.2 EA-Reliable

Table 16 illustrates an *EA-Reliable* failure case concerning the botanical nature of bananas. The KA incorrectly labels the claim as FALSE, succumbing to a deeply rooted colloquial stereotype that bananas grow on woody trees. Conversely, the EA retrieves precise scientific facts explaining that the banana plant is a large herb and that its apparent trunk is actually a pseudostem lacking true woody tissue. In this case, the AA exhibits an over-defensive bias

Dataset	Model/Setting	# FN Errors	# FP Errors	Ratio FN %	Ratio FP %
Snopes	Qwen (KA)	720	106	87.17%	12.83%
	Qwen (EA)	251	284	46.92%	53.08%
	Llama (KA)	578	274	67.84%	32.16%
	Llama (EA)	330	178	64.96%	35.04%
PolitiFact	Qwen (KA)	1,226	35	97.22%	2.78%
	Qwen (EA)	601	278	68.37%	31.63%
	Llama (KA)	1,112	98	91.90%	8.10%
	Llama (EA)	625	218	74.14%	25.86%

Table 13: Granular error distribution for Qwen2.5-7B-Instruct (Qwen) and Llama3.1-8B-Instruct (Llama) models, comparing KA and EA. This table analyzes incorrect predictions: **FN** (False Negatives) are truthful claims labeled FALSE, and **FP** (False Positives) are false claims labeled TRUE.

and selects TrustKA. It dismisses the precise botanical definition provided by the EA as anomalous or highly pedantic noise, blindly preferring the KA’s intuitive common-sense stereotype. This case exposes a limitation of the framework: it remains overly defensive toward valid external evidence that challenges entrenched beliefs.

L.3 Both-Wrong

Table 17 examines a *Both-Wrong* failure case regarding a pseudoscientific medical belief. In this case, both specialized agents initially produce an incorrect TRUE verdict: the KA suffers from flawed causal reasoning regarding heat and viral replication, while the EA confuses comforting symptomatic relief with a factual biological cure by retrieving unverified health blogs. Although their high-confidence agreement triggers the Reflective Rejection mechanism, this intervention ultimately fails to break the error loop. During re-evaluation, each agent generates sophisticated rationalizations to lock in the error: the revised KA constructs a plausible-sounding but flawed explanation regarding viral denaturation, while the revised EA treats cultural consensus as factual proof of virus eradication. This failure illustrates that even with active reflective mechanisms, the framework can still lock in entrenched patterns of erroneous groupthink.

Stage	Prompt Template and Instructions
Knowledge Agent	System: You are an expert fact-checker relying solely on internal knowledge. User: For the claim: "{claim}" 1. First, select a reasoning strategy: – FAST : Deliver your verdict immediately for well-known facts. – STEP : Articulate a step-by-step logical analysis for complex assertions. 2. Then, output your response in the following exact format: Strategy: <FAST or STEP> Judgment: <TRUE or FALSE> Confidence: <a float in [0.0, 1.0]>
Evidence Agent	System: You are an expert evidence-based fact-checker. User: Given the claim: "{claim}" and evidences: {evidence_list} 1. First, select a reasoning strategy: – PIECE : Verify the claim against each snippet, checking literal isomorphism. – SUM : Assess the claim based on a synthesized summary of evidence. 2. Then, output your response in the following exact format: Strategy: <PIECE or SUM> Judgment: <TRUE or FALSE> Confidence: <a float in [0.0, 1.0]>
Conflict Perception	System: You are a precise meta-reasoning classifier. User: Evaluate two judgments for the claim below: Claim: "{claim}" Evidences: {evidence_list} Judgments: – KA: "{ka_pred}" [internal knowledge] – EA: "{ea_pred}" [external evidence] Select ONE type: • Type 1 : Agree and correct → accept consensus. • Type 2 : Disagree, trust KA more → choose KA. • Type 3 : Disagree, trust EA more → choose EA. • Type 4 : Agree but likely wrong → trigger Reflective Rejection . Respond ONLY with: "Type 1", "Type 2", "Type 3", or "Type 4".
Reflective Rejection	System: You are a meta-cognitive adjudicator tasked with synthesizing multiple fact-checking perspectives to reach a robust final verdict. (Type 4 only) User: Below are four independent analyses of the same claim: [KA-FAST], [KA-STEP], [EA-PIECE], [EA-SUM] 1. First, your goal is not to average opinions, but to – Identify inconsistencies between internal knowledge and external evidence. – Use strategy flips as diagnostic signals for reasoning instability. – Mitigate overreliance on literal evidence or dismissal of internal knowledge. – Reconcile the logic to assess if evidence insufficiency warrants rejection. 2. Then, output your final decision in this exact format: Judgment: <TRUE or FALSE>

Table 14: Standardized prompt templates for the modular components. The templates define the functional transition from parametric reasoning (KA) to evidence grounding (EA) and meta-cognitive adjudication (AA).

Claim	"The tech company banned all remote work globally in early 2018."
Knowledge Agent (KA)	Verdict: FALSE <i>[Summary]</i> According to stable internal parametric knowledge regarding corporate policies up to early 2018, the company expanded its hybrid work modes rather than instituting a total global ban.
Evidence Agent (EA)	Verdict: TRUE <i>[Summary]</i> The EA retrieves sensational social media snippets that react to a leaked internal memo discussing potential localized restrictions, misinterpreting it as an actual global ban.
Adjudication Agent (AA)	Mode: TrustEA <i>[Summary]</i> The AA over-relies on the retrieved external text, wrongly assuming that the prominent and emotionally charged social media buzz outweighs the stable parametric knowledge of the KA.
Final Verdict	TRUE (Ground Label: FALSE)

Table 15: *KA-Reliable* failure case involving a misjudgment caused by noisy information. The KA holds accurate knowledge, but the EA is misled by noise; the AA over-relies on external evidence, leading to an incorrect verdict.

Claim	“Bananas grow on giant herbaceous plants, not trees.”
Knowledge Agent (KA)	Verdict: FALSE [Summary] The KA succumbs to a common colloquial stereotype that bananas grow on woody trees with standard arborescent structures.
Evidence Agent (EA)	Verdict: TRUE [Summary] The banana plant is a large herbaceous plant. What appears to be a trunk is actually a pseudostem made of tightly packed leaf bases, lacking true woody tissue.
Adjudication Agent (AA)	Mode: TrustKA [Summary] It dismisses the precise botanical definition provided by the EA as anomalous or highly pedantic noise, blindly preferring the intuitive common-sense stereotype held by the KA.
Final Verdict	FALSE (Ground Label: TRUE)

Table 16: *EA-Reliable* failure case involving deeply rooted misconceptions. The KA falls for a colloquial stereotype while the EA provides precise evidence; the AA exhibits a defensive bias, refusing to update internal beliefs.

Claim	“Drinking hot water during a fever cures the underlying infection by boiling the pathogens.”
Knowledge Agent (KA)	Verdict: TRUE [Summary] The KA suffers from flawed causal reasoning: Heat kills viruses, so consuming hot liquids directly assists the immune system in pasteurizing the infection.
Evidence Agent (EA)	Verdict: TRUE [Summary] The EA retrieves unverified health blogs stating that hot water flushes out and destroys viruses, confusing comforting symptomatic relief with a factual biological cure.
Adjudication Agent (AA)	Mode: Reflective Rejection Instructs the KA and EA to flip their reasoning strategies.
Revised KA	[Summary] Higher body temperatures inhibit viral replication, thereby implying that external heat sources like hot water can directly accelerate viral denaturation. → Verdict: TRUE
Revised EA	[Summary] It finds that dozens of cultural health articles endorse this practice, mistakenly treating this correlation as validation of biological antiviral efficacy. → Verdict: TRUE
Final Verdict	TRUE (Ground Label: FALSE)

Table 17: *Both-Wrong* case involving pseudoscientific consensus. Although the AA triggers Reflective Rejection, both agents suffer from confirmation bias during re-evaluation, generating rationalizations to entrench the error.