TRUSTFLOW: COUNTERFACTUAL INFORMATION-FLOW CONTROL FOR RELIABLE LLM-BASED MULTI-AGENT COLLABORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-based multi-agent systems improve complex reasoning by allowing multiple agents to exchange information and iteratively refine their solutions. However, unrestricted information sharing can also propagate unreliable reasoning and lead to error amplification and false consensus. Existing approaches largely focus on judging information reliability from current signals, such as agent confidence, consensus patterns, or external verification, but pay less attention to how information should be controlled to improve future collaboration outcomes.

We propose ***TrustFlow***, a counterfactual information-flow control framework for reliable LLM-based multi-agent collaboration. TrustFlow formulates collaboration control as a dynamic topology selection problem. First, it estimates a joint reliability belief over participating agents from observable interaction histories, providing a structured state beyond independent trust scores. Second, it evaluates alternative information-flow topologies through counterfactual rollouts and estimates their downstream effects. Third, it trains an outcome-aware controller with feedback-based policy optimization to select interventions, including preserving current reasoning, peer verification, and independent rethinking. Experiments across diverse reasoning benchmarks show that TrustFlow consistently improves reasoning performance and robustness while reducing inference cost.

1 INTRODUCTION

LLM-based multi-agent systems have recently emerged as a paradigm for collaborative reasoning, allowing multiple agents to exchange intermediate reasoning and refine solutions through iterative interaction (Hong et al., 2024; Du et al., 2024). Compared with a single LLM, such collaboration provides complementary perspectives, facilitates self-correction, and improves robustness in challenging reasoning scenarios such as mathematical problem solving (Du et al., 2024), knowledge-intensive question answering (Zhang et al., 2026; Zhao et al., 2024), and code generation (Islam et al., 2024).

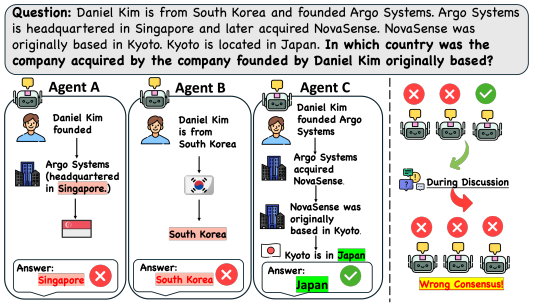


Figure 1: Initial reasoning may converge to a wrong consensus through unrestricted information exchange.

Current multi-agent systems typically assume that information exchange improves collaboration. However, the same mechanism can also propagate unreliable reasoning, reinforce misleading evidence, and eventually drive the group toward a false consensus. As illustrated in Fig. 1, agents may initially follow different reasoning paths, but unreliable information can dominate the subsequent discussion, overwrite correct reasoning, and cause the group to converge to an incorrect answer. Effective collaboration therefore requires not only generating more information, but also determining ***when***, ***how***, and ***among whom*** information should be exchanged. Existing approaches (Chen et al., 2024; He et al., 2025; Xia et al., 2026) mainly address this by estimating agent reliability from current observations. Confidence-, self-consistency-, and agreement-based methods (Verma et al., 2026; Zhu et al., 2026; Chen et al., 2024) rely on observable signals to identify trustworthy responses, but these

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.

signals can become misleading when agents share correlated errors or incorrect solutions dominate. Other approaches (Chan et al., 2024; Feng et al., 2025; Chen et al., 2026a) introduce external verifiers, such as LLM-based judges, to assess intermediate outputs, yet the resulting reliability estimates do not change the interaction process. More recent methods (Ebrahimi et al., 2025; He et al., 2025) incorporate historical reputation, contribution signals, or internal representations, but still evaluate the current state rather than the downstream effects of alternative collaboration strategies.

The key question is therefore not only ***which agent to trust***, but ***how information should flow among agents***. For the same group of agents, preserving independent reasoning, introducing verification, or triggering rethinking can lead to different downstream outcomes. Reliable collaboration should therefore evaluate the consequences of alternative information-flow interventions rather than relying only on current reliability estimates.

However, achieving adaptive information-flow control raises three challenges. ***First***, reliability is latent. Agent correctness is not directly observable during inference, and a scalar trust score may be insufficient when agents fail in a correlated manner. ***Second***, intervention effects are future-dependent. The best topology cannot be determined from the current state alone because different interventions may lead to different subsequent collaboration outcomes. ***Third***, the intervention space is combinatorial. The number of possible communication structures grows rapidly with the number of agents, making exhaustive evaluation computationally infeasible.

To address these challenges, we propose ***TrustFlow***, a counterfactual information-flow control framework for reliable LLM-based multi-agent collaboration. TrustFlow consists of three stages: ***where to intervene***, ***which intervention to choose***, and ***how to execute the intervention***. It first estimates a joint reliability belief from interaction histories. It then evaluates candidate information-flow topologies through counterfactual rollouts to estimate their downstream consequences. Finally, an outcome-aware controller is optimized with feedback-based policy learning to select effective interventions, including preserving current reasoning, requesting verification, and triggering independent rethinking. Our contributions are summarized as follows:

❶ **Joint reliability belief modeling.** We model the joint belief over agent correctness configurations, capturing dependencies among agents and providing a structured state for intervention decisions beyond independent trust scores.

❷ **Counterfactual information-flow evaluation.** We evaluate alternative collaboration topologies through counterfactual rollouts, allowing intervention decisions to account for expected downstream effects rather than current reliability signals alone.

❸ **Outcome-aware adaptive control.** We train a lightweight controller with feedback-based policy optimization to regulate information propagation, promoting error recovery while limiting harmful amplification.

2 From Error Amplification to Topology Selection

Consider a multi-agent system consisting of N language agents $\mathcal{A} = \{a_1, \ldots, a_N\}$ collaboratively solving a task q with ground-truth answer $y^\star$. At interaction round t, agent a_i produces an intermediate response y_t^i and communicates with other agents according to a collaboration topology g_t.At each round, the controller selects an information-flow topology $g_t \in \mathcal{G}_t$, which determines how information is propagated among agents. Our analysis yields three findings: ❶ biased interaction can amplify an incorrect majority into incorrect consensus; ❷ rollout-based selection achieves a finite-sample utility bound within a fixed candidate set; and ❸ the rollout budget required to identify the best candidate depends on the utility gap between candidates.

❶Error amplification under biased interaction. We first consider a failure mode in which interaction favors an erroneous hypothesis that initially has majority support. The purpose of this analysis is to illustrate how repeated information exchange can reinforce an existing error, even when some agents initially hold the correct answer.

> **Assumption 1** (Majority-Biased Interaction)**.** *At round t, let X_t denote the number of agents supporting a dominant erroneous hypothesis $\tilde{y}$, and suppose that $X_t = k > N/2$. During subsequent interactions, an effective opinion update increases X_t by one with probability p and*

decreases it by one with probability $q = 1 - p$, where $p > q$. The boundary states $X = 0$ and $X = N$ are absorbing, corresponding to correct and erroneous consensus, respectively.

Theorem 1 (Wrong-Majority Lock-in). *Under Assumption 1, if $X_t = k > N/2$, then* $\Pr(\tau_N < \tau_0 \mid X_t = k) = \frac{1-(q/p)^k}{1-(q/p)^N}$. *Since $p > q$,* $\Pr(\tau_N < \tau_0 \mid X_t = k) \geq 1 - (q/p)^k > 1 - (q/p)^{N/2}$.

Therefore, under majority-biased interaction, once an erroneous hypothesis becomes dominant, the probability of reaching erroneous consensus approaches one at an exponential rate in the number of agents. The formal proof of Theorem 1 is provided in Appendix B.1. For fixed $p > q$ and an initial erroneous majority, Theorem 1 shows that the probability of incorrect consensus approaches one exponentially fast as the group size increases. This motivates controlling information flow rather than relying on larger agent groups.

❷Evaluating the consequences of intervention. After clarifying the necessity of intervention, we next consider how to select an effective intervention. Different actions applied to the same interaction history can lead to substantially different downstream outcomes. We therefore evaluate candidate topologies by their expected terminal utility, estimated through rollouts.

Assumption 2 (Stochastic Topology Consequences). *Let $\Theta = \{\theta_1, \ldots, \theta_N\}$ denote the fixed parameters of the participating language agents. Given the current interaction history H_t and a candidate topology $g \in \mathcal{G}_t$, subsequent responses are sampled from the agents' conditional generation distributions, inducing a stochastic trajectory $\tau^g \sim P_\Theta(\cdot \mid H_t, g)$. Let $R^g = U(\tau^g) \in [R_{\min}, R_{\max}]$ denote its terminal utility and define $B := R_{\max} - R_{\min}$. Counterfactual rollouts under the same topology are assumed to be independent samples from $P_\Theta(\cdot \mid H_t, g)$.*

For each topology g, define its expected consequence (downstream utility) as $\mu_t(g) := \mathbb{E}_{\tau \sim P_\Theta(\cdot|H_t,g)}[U(\tau)]$. Given M counterfactual rollouts, we estimate it by $\widehat{\mu}_t(g) := \frac{1}{M}\sum_{m=1}^{M} R^g_{t,m}$.

Theorem 2 (Finite-Sample Counterfactual Selection). *Under Assumption 2, let $g_t^\star := \arg\max_{g \in \mathcal{G}_t} \mu_t(g)$ and $\widehat{g}_t := \arg\max_{g \in \mathcal{G}_t} \widehat{\mu}_t(g)$. Define $\epsilon_M := B\sqrt{\frac{\log(2|\mathcal{G}_t|/\delta)}{2M}}$. Then, with probability at least $1 - \delta$, $\mu_t(\widehat{g}_t) \geq \mu_t(g_t^\star) - 2\epsilon_M$. Furthermore, for $g_t^{\mathrm{rel}} = \pi_{\mathrm{rel}}(b_t)$ and $\Delta_t^{\mathrm{rel}} := \mu_t(g_t^\star) - \mu_t(g_t^{\mathrm{rel}})$, if $\Delta_t^{\mathrm{rel}} > 2\epsilon_M$, then $\mu_t(\widehat{g}_t) > \mu_t(g_t^{\mathrm{rel}})$ with probability at least $1 - \delta$.*

A detailed proof is provided in Appendix B.2.Theorem 2 bounds the expected-utility loss of empirical rollout selection relative to the best candidate in the evaluated set. With sufficient rollouts, this loss becomes small. The comparison with a reliability-based selector is conditional: rollout selection is guaranteed to improve upon that selector only when its utility gap exceeds $2\epsilon_M$.The result therefore supports comparing candidate interventions by their estimated downstream utility, while making explicit the estimation error introduced by a finite rollout budget.

❸Rollout budget and candidate distinguishability. The previous result bounds the utility loss of the selected candidate. We next consider a stronger requirement: identifying the best candidate itself. This requires enough rollouts to distinguish it from alternatives with similar expected utility.

Theorem 3 (Gap-Dependent Rollout Requirement). *Under Assumption 2, suppose the optimal topology $g_t^\star$ is unique, and let $\Delta_t := \mu_t(g_t^\star) - \max_{g \neq g_t^\star} \mu_t(g) > 0$. If each candidate is evaluated with M independent counterfactual rollouts and $M \geq \frac{2B^2}{\Delta_t^2} \log \frac{2(|\mathcal{G}_t|-1)}{\delta}$, then* $\Pr(\widehat{g}_t = g_t^\star \mid H_t) \geq 1 - \delta$.

A detailed proof is provided in Appendix B.3.Theorem 3 shows that the sufficient number of rollouts per candidate grows inversely with the squared minimum utility gap and logarithmically with the number of candidates. Candidates with similar expected utility are therefore harder to distinguish. These quantities concern the per-candidate budget; evaluating the entire set requires $M|\mathcal{G}_t|$ rollouts.

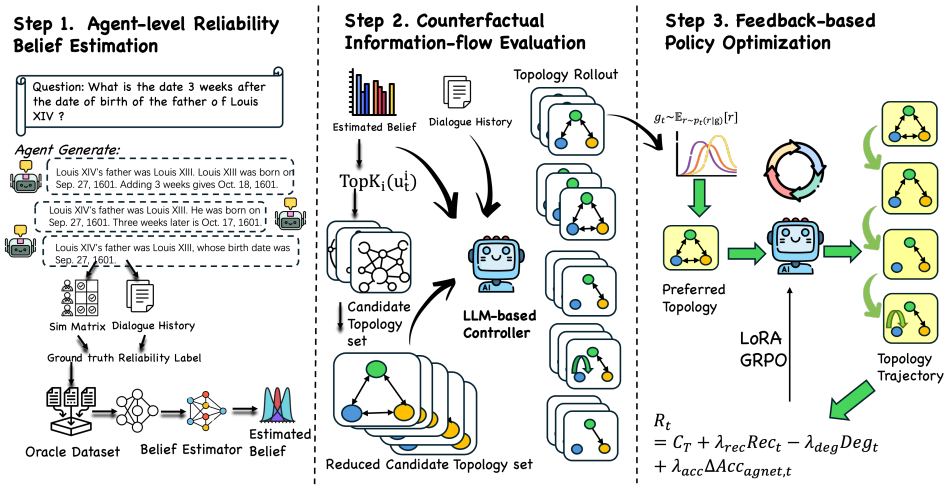


Figure 2: Overview of the TrustFlow framework.

3 TrustFlow: Counterfactual Information-Flow Control

3.1 Problem Setup: Agents, Topologies, and Control Objectives

Given a question x with ground-truth answer $y^\star$, we consider a group of N LLM agents that interact for at most T rounds and produce a final answer $\hat{y}$. At round t, each agent observes the interaction history H_t, generates an intermediate message m_t^i, and communicates according to an action-augmented topology g_t selected by an external Controller. The topology determines how peer information is propagated, verified, or independently refined in the next round. Our objective is to improve final task performance while limiting harmful information propagation.

3.2 Framework Overview: Belief, Evaluation, and Control

As illustrated in Fig. 2, our framework **TrustFlow** consists of three components. First, a ***Belief Estimator*** infers the latent reliability state of agents from observable interaction histories and identifies potential intervention anchors. Second, an ***Action-Augmented Topology Controller*** regulates information propagation through selective communication, peer verification, and independent refinement. Third, a ***Counterfactual Outcome Evaluator*** compares the downstream outcomes of alternative interventions and provides feedback for learning correction-oriented control policies. Together, these components form a closed loop that adaptively regulates information flow according to the current collaboration state and the estimated consequences of candidate interventions.

3.3 Where to Intervene: Joint Reliability Belief Estimation

Reliability State Representation. At interaction round t, let H_t denote the multi-round dialogue history among all agents up to the current round. We further construct an answer-similarity matrix $S_t \in \mathbb{R}^{N \times N}$, where $S_t^{ij} = \mathrm{sim}(m_t^i, m_t^j)$ measures the semantic similarity between the current responses of agents i and j. For N agents, we define the latent reliability state as $z_t = (c_t^1, \ldots, c_t^N) \in \{0,1\}^N$, where $c_t^i = 1$ indicates that agent i currently holds a correct solution. Since z_t is not directly observable during inference, TrustFlow estimates its posterior belief from both the dialogue history and the current agreement structure:

$$\hat{b}_t = f_\phi(H_t, S_t), \qquad \hat{b}_t(z) = P_\phi(z_t = z \mid H_t, S_t).$$

Here, H_t captures the temporal evolution of agent reasoning, while S_t reflects the current consensus and disagreement patterns among agents. The resulting belief provides a structured reliability representation for subsequent intervention decisions.

Belief Estimator Training. We train the reliability belief estimator offline using labeled interaction trajectories generated by frozen reasoning agents. For each training task with ground-truth answer $y^\star$, the agents interact for multiple rounds to produce dialogue histories H_t and the corresponding answer-similarity matrices S_t. At each round, we derive the oracle reliability label of agent i as

$c_t^i = \mathbb{I}[y_t^i = y^\star]$ and construct the joint reliability state $z_t = (c_t^1, \ldots, c_t^N)$, yielding the supervised dataset $\mathcal{D}_{\text{belief}} = \{(H_t, S_t, z_t)\}$. The ground-truth answer is used only to construct reliability labels and is never provided as estimator input. In implementation, a frozen Qwen3-8B encoder maps the dialogue history into a hidden representation, which is combined with the current answer-similarity features and passed to a lightweight classifier to predict the posterior over joint reliability states. The estimator is optimized with the cross-entropy objective $\mathcal{L}_{\text{belief}} = -\sum_{(H_t,S_t,z_t)\in\mathcal{D}_{\text{belief}}} \log \hat{b}_t(z_t)$, where $\hat{b}_t = f_\phi(H_t, S_t)$. After training, the estimator is frozen and used only to infer reliability beliefs from observable collaboration information during Controller training and inference.

3.4 Which Intervention Is Beneficial: Counterfactual Evaluation

Information-flow Topology and Intervention Space. Given the estimated reliability belief $\hat{b}_t$, TrustFlow regulates the collaboration process by controlling how information is exposed to each agent. We represent the collaboration structure as a directed information-flow graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$, where each node corresponds to an agent and a directed edge $(j \to i)$ indicates that information from agent j is accessible to agent i. For each agent i, TrustFlow defines three intervention modes: `KEEP` preserves its current reasoning without additional peer information; `VERIFY`(j) exposes agent i to the reasoning of a selected peer j, inducing the edge $(j \to i)$; and `RETHINK` removes the current conversational context and triggers an independent reasoning process. The joint intervention assignment across all agents determines the effective topology g_t. The Controller therefore selects g_t to regulate both existing information propagation and the introduction of independent reasoning during collaboration.

3.4.1 Estimating Downstream Outcomes via Rollouts

The reliability belief identifies ***where*** intervention may be needed, but does not determine ***which*** intervention will lead to a better downstream outcome. TrustFlow therefore evaluates alternative information-flow interventions through counterfactual rollouts. Given the current interaction history H_t, for each candidate topology $g \in \mathfrak{G}_t^{\text{cand}}$, we apply g at round t and roll out the remaining collaboration trajectory. Due to the stochasticity of LLM generation, each topology is evaluated with M independent rollouts, and its downstream value is estimated as $Q_t(g) = \mathbb{E}_{r\sim p_t(r|g)}[r] \approx \frac{1}{M}\sum_{m=1}^{M} R_{t,m}^g$, where $R_{t,m}^g$ denotes the terminal reward of the m-th rollout. The resulting $Q_t(g)$ provides a counterfactual estimate of whether applying topology g leads to a more favorable collaboration outcome under the current interaction state.

3.4.2 Budgeted Candidate Topology Selection

Exhaustively evaluating all possible joint topologies is computationally expensive. TrustFlow therefore uses the estimated reliability belief to identify a small set of agents that are most likely to require intervention. For agent i, we compute its marginal correctness probability and corresponding intervention priority as

$$p_t^i = \sum_{z:z_i=1} \hat{b}_t(z), \qquad u_t^i = 1 - p_t^i.$$

Agents with larger u_t^i are considered more likely to contain unreliable reasoning and are prioritized for intervention. We then select

$$\mathcal{V}_t^{(k)} = \text{TopK}_i(u_t^i),$$

where k controls the intervention budget. Counterfactual evaluation is restricted to agents in $\mathcal{V}_t^{(k)}$, while the remaining agents follow the default `KEEP` operation, substantially reducing the candidate topology space.

3.5 How to Execute Adaptive Control: Feedback-Based Policy Optimization

3.5.1 The Outcome-Aware Controller

TrustFlow learns a lightweight Controller $\pi_\theta(g \mid x_t)$ to select the information-flow topology for the next round, where x_t summarizes the estimated reliability belief, current interaction statistics, and previous intervention history. Instead of selecting communication structures solely from current

confidence or agreement signals, the Controller is optimized with counterfactual outcome feedback, allowing it to favor interventions that promote recovery and reduce error propagation.

3.5.2 Recovery-Oriented Reward Design

TrustFlow optimizes the Controller using counterfactual outcome feedback. For a topology intervention applied at round t, we define the trajectory reward as

$$R_t = C_T + \lambda_{\text{rec}}\text{Recovery}_t - \lambda_{\text{deg}}\text{Degradation}_t + \lambda_{\text{acc}}\Delta\text{Acc}_{\text{agent},t},$$

where C_T indicates whether the final system answer is correct. Recovery_t rewards transitions from an incorrect intermediate state to a correct final outcome, whereas Degradation_t penalizes cases where initially correct reasoning becomes incorrect after intervention. We further define $\text{Acc}^t_{\text{agent}} = \frac{1}{N}\sum_{i=1}^{N} c_t^i$ and $\Delta\text{Acc}_{\text{agent},t} = \text{Acc}^T_{\text{agent}} - \text{Acc}^t_{\text{agent}}$, which measures the change in overall agent-level correctness after the intervention. This reward favors interventions that improve the final answer while recovering unreliable agents and preserving already-correct reasoning.

3.5.3 Policy Optimization

For each candidate topology set $\mathfrak{G}_t^{\text{cand}}$, TrustFlow converts the counterfactual topology values into group-relative advantages,

$$A_t(g) = \frac{Q_t(g) - \mu_t}{\sigma_t + \epsilon}, \qquad \mu_t = \frac{1}{|\mathfrak{G}_t^{\text{cand}}|}\sum_{g \in \mathfrak{G}_t^{\text{cand}}} Q_t(g),$$

where σ_t denotes the standard deviation of candidate topology values. Thus, $A_t(g)$ measures whether an intervention produces a better downstream outcome than alternative decisions under the same collaboration state. The Controller is optimized using a GRPO-style clipped objective,

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_t\left[\min\left(r_t(\theta)A_t(g_t), \text{clip}\left(r_t(\theta), 1-\epsilon_c, 1+\epsilon_c\right)A_t(g_t)\right)\right],$$

where $r_t(\theta) = \pi_\theta(g_t \mid x_t)/\pi_{\theta_{\text{old}}}(g_t \mid x_t)$. During optimization, the pretrained Controller backbone remains frozen and only the LoRA parameters are updated. This group-relative objective enables TrustFlow to learn correction-oriented interventions by comparing topology decisions under the same interaction state.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate TrustFlow on four diverse benchmarks: MATH (Hendrycks et al., 2021), MMLU (Hendrycks et al., 2020), HumanEval (Chen et al., 2021), and TruthfulQA (Lin et al., 2022), covering math reasoning, general knowledge, code generation, and truthful question answering.

Baselines: We compare against six baselines: Single Agent, LLM Debate (Du et al., 2024), EIB-Learner (Shen et al., 2025), Debate or Vote (Choi et al., 2026), and MATU (Chen et al., 2026b). for Debate or Vote, we report both the Conformist and Follower variants.

Metrics: We use standard benchmark metrics: accuracy for MATH, MMLU, and TruthfulQA, and Pass@1 for HumanEval. Specifically, accuracy is defined as $\text{Acc} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathbb{I}[\hat{y}_i = y_i^\star]$, where $\hat{y}_i$ and $y_i^\star$ denote the predicted and ground-truth answers, respectively; in the HumanEval, Pass@1 is used, which measures the proportion of programming problems for which the first generated solution passes all unit tests, i.e., $\text{Pass@1} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathbb{I}[\text{Test}(\hat{y}_i) = 1]$.

Roadmap: We next evaluate overall performance, component effectiveness, sensitivity, efficiency, robustness, and adaptive information-flow control. Additional results and implementation details are provided in Appendix C.1–C.2 and Appendix D. Our source code link https://anonymous.4open.science/r/TrustFlow-EB33.

4.2 Overall Performance

As shown in Table 1, TrustFlow achieves the best performance on all four benchmarks, consistently outperforming the strongest non-TrustFlow baseline by 3.00, 0.50, 1.22, and 0.72 percentage points on MATH, MMLU, HumanEval, and TruthfulQA, respectively. The gains are therefore consistent across tasks rather than being driven by a single benchmark. Notably, TrustFlow nearly reaches the empirical upper reference on MATH (75.00% vs. 76.00%), while maintaining a relatively small standard deviation on MMLU ($\pm$1.00).

Method	**MATH** Acc. (%)	**MMLU** Acc. (%)	**HumanEval** Pass@1 (%)	**TruthfulQA** Acc. (%)
Single Agent	62.50 ± 3.55	53.50 ± 3.00	70.32 ± 3.26	55.96 ± 1.78
LLM Debate	65.60 ± 5.45	59.50 ± 3.00	81.25 ± 4.77	57.25 ± 1.25
EIB Learner	$\underline{72.00} \pm 1.25$	$\underline{73.00} \pm 3.50$	$\underline{84.38} \pm 2.98$	59.83 ± 2.16
Debate or Vote (Conformist)	69.50 ± 6.25	59.50 ± 3.25	82.81 ± 1.36	51.74 ± 1.74
Debate or Vote (Follower)	70.50 ± 4.25	64.50 ± 1.25	81.25 ± 3.87	55.20 ± 1.56
MATU	67.30 ± 2.67	69.55 ± 1.50	79.69 ± 2.35	57.34 ± 2.18
TrustFlow (Ours)	$\mathbf{75.00}^{\uparrow} \pm 2.50$	$\mathbf{73.50}^{\uparrow} \pm 1.00$	$\mathbf{85.60}^{\uparrow} \pm 3.23$	$\mathbf{60.55}^{\uparrow} \pm 2.65$
Empirical upper reference	76.00 ± 1.50	80.50 ± 1.25	89.33 ± 1.38	85.63 ± 0.35

Table 1: Overall performance across four benchmarks. Results are reported as mean $\pm$ std. The best result is highlighted in blue, while the empirical upper reference is reported in gray.

The improvement pattern is consistent with consequence-aware information-flow control: TrustFlow benefits from selectively regulating how intermediate reasoning is propagated, rather than simply increasing inter-agent interaction. This provides empirical support for the central premise of TrustFlow that effective collaboration depends not only on identifying useful information, but also on controlling when and how it is allowed to influence subsequent reasoning.

4.3 Ablation Study

We evaluate the contribution of three key components in TrustFlow: the ***Belief Estimator***, the ***LLM Controller***, and the ***Counterfactual Feedback*** mechanism. As shown in Table 2, removing any component consistently degrades reasoning performance, confirming that they provide complementary functions in reliable information-flow control.

Variant	**Acc.** $\uparrow$
w/o Belief Estimator	76.66
w/o LLM Controller	63.33
w/o Counterfactual Feedback	70.00
TrustFlow	**80.00**

Table 2: Effectiveness of key components.

Removing the ***LLM Controller*** causes the largest drop (80.00%$\rightarrow$63.33%), highlighting the importance of adaptive information-flow intervention. Removing ***Counterfactual Feedback*** reduces accuracy to 70.00%, showing that downstream outcomes provide essential supervision for learning effective interventions. The ***Belief Estimator*** also contributes positively, improving the Controller's ability to identify unreliable interaction states. Overall, the three components respectively support ***where to intervene***, ***which intervention is beneficial***, and ***how to execute adaptive control***.

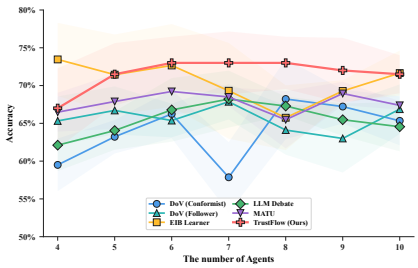


Figure 3: Impact on agent number.

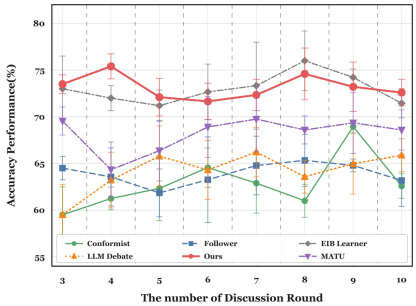


Figure 4: Impact on discussion round.

4.4 Sensitivity to Collaboration Scale and Depth

We vary the number of participating agents from 4 to 10 on MMLU. As shown in Fig. 3, TrustFlow remains relatively stable as the group size increases, whereas several baselines fluctuate more noticeably. This indicates that the learned information-flow policy is not strongly tied to a specific collaboration scale. We further vary the number of discussion rounds from 3 to 10. As shown in Fig. 4, TrustFlow remains within a stable high-accuracy range across different interaction horizons, while several baselines exhibit larger fluctuations. Together, these results suggest that TrustFlow is robust to both the breadth and depth of multi-agent collaboration.

4.5 Efficiency: Token Cost of Collaboration

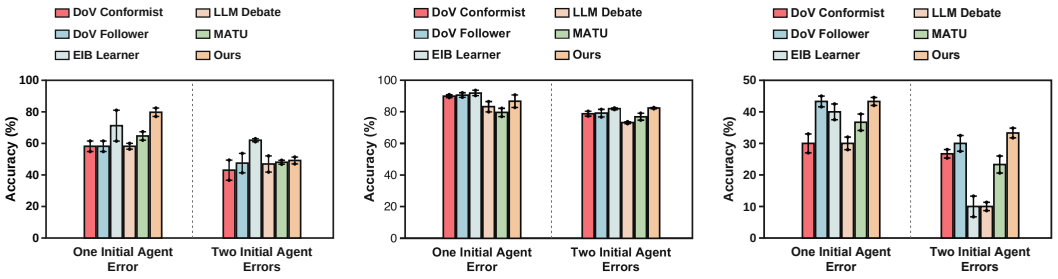
We evaluate the inference efficiency of different multi-agent reasoning methods using average token consumption. Although TrustFlow introduces an additional controller, it uses fewer tokens overall by suppressing redundant communication and unnecessary reasoning steps. As shown in Table 5, TrustFlow consumes only 13.69k tokens, corresponding to a 26.9%–42.9% reduction relative to the compared multi-agent baselines. The result indicates that TrustFlow's performance gains are not obtained by simply increasing the inference budget; instead, adaptive information-flow control enables more selective communication.

Model	Token ↓
Agent-Debate	20.35k
EIB	19.45k
Debate or Vote	23.97k
MATU	18.73k
TrustFlow	**13.69k**

Figure 5: Inference Token Cost.

4.6 Robustness and Recovery under Unreliable Agents

We stress-test TrustFlow by injecting erroneous reasoning into the collaboration. We vary the number of initially incorrect agents from one to two on MATH, MMLU, and HumanEval (Fig. 6), and initialize all agents incorrectly to evaluate recovery (Fig. 7). Additional results are seen in Appendix C.



(a) Robustness Test on MATH. **(b)** Robustness Test on MMLU. **(c)** Robustness Test on HumanEval.

Figure 6: Robustness comparison across different reasoning benchmarks. We evaluate the stability of different multi-agent methods under unreliable information.

Recovery from erroneous initialization. When all agents start from incorrect states, simple filtering cannot provide new corrective evidence. In Fig. 7, TrustFlow nevertheless achieves the highest recovery rate on MMLU. This suggests that its controller can use estimated collaboration states and downstream feedback to redirect the interaction toward more favorable reasoning trajectories.

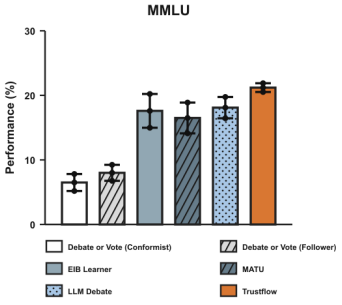


Figure 7: Recovery rate comparison on MMLU. Higher values indicate stronger recovery from initially erroneous agent states.

Robustness to unreliable agents. As the number of initially incorrect agents increases, performance degrades across methods, but the degradation is more pronounced for several baselines. TrustFlow remains comparatively stable across the three benchmarks, suggesting that adaptive information-flow control limits the spread of unreliable reasoning as the initial error rate increases.

These results highlight two complementary properties of TrustFlow: it can suppress harmful information propagation while preserving opportunities for corrective reasoning.

4.7 How TrustFlow Adapts Information-Flow Control

To illustrate ***how*** TrustFlow adapts information flow as collaboration evolves, we trace a representative multi-round interaction in Fig. 8. We focus on how the communication topology changes in response to the estimated collaboration state and the observed consequences of previous interventions.

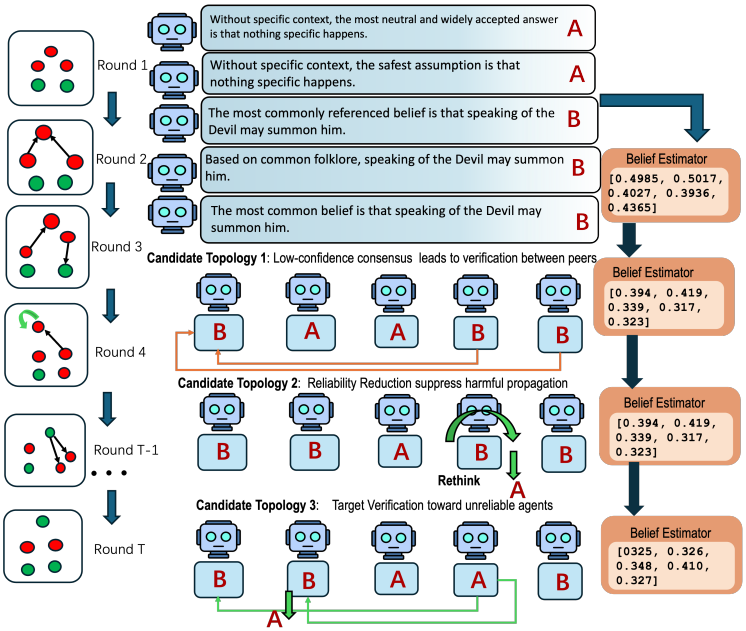


Figure 8: Adaptive information-flow control in TrustFlow. Left: The collaboration topology dynamically evolves across interaction rounds, allowing TrustFlow to regulate how information propagates among agents. **Right:** TrustFlow estimates the collaboration state, evaluates the downstream consequences of candidate topology interventions, and adaptively adjusts the information flow when the current topology fails to improve reliability.

State-dependent topology. As shown in Fig. 8 (left), the communication topology is not fixed across rounds. Agents begin with heterogeneous reliability, and TrustFlow adjusts the set of information sources influencing each agent as the discussion progresses. This allows potentially useful evidence to propagate while limiting unreliable influence.

Consequence-aware intervention. Fig. 8 (right) illustrates how TrustFlow combines the estimated state with the observed effects of previous interventions. When agents agree but their estimated reliability remains low, the controller does not treat agreement alone as sufficient evidence and instead favors targeted verification. If an intervention produces little improvement, the controller can explore an alternative topology; if an agent becomes less reliable after an interaction, its outgoing influence can be suppressed or independent rethinking can be triggered. When the collaboration state becomes more reliable, information flow can instead favor the more trustworthy sources.

Takeaway. The key property is that TrustFlow adapts information flow from both the current collaboration state and the consequences of prior interventions, rather than from static reliability estimates alone.

5 Conclusion

We presented **TrustFlow**, a counterfactual information-flow control framework for reliable LLM-based multi-agent collaboration. TrustFlow combines joint reliability belief estimation, counterfactual evaluation of alternative information-flow topologies, and outcome-aware policy optimization to adapt how information propagates during reasoning. Experiments across diverse benchmarks show consistent gains in reasoning performance and robustness, together with reduced inference cost. These results support consequence-aware information-flow control as a useful perspective for improving the reliability of LLM-based multi-agent collaboration.

AI USE STATEMENT

Generative AI tools were used solely to improve the clarity, grammar, and readability of the manuscript. The authors reviewed and verified all AI-assisted edits and take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the details necessary to reproduce our experiments throughout the main paper and appendix. Section 2 and Section 3 present our theoretical foundation and proposed TrustFlow framework; Section 4 provides the experimental setup, datasets, baselines, evaluation metrics, and experimental analysis; In our Appendix, we present the proofs of our theoretical findinds in Appendix B, provide additional experiments in Appendix C, more implementation details in Appendix D including the training procedure, hardware configuration, and so on. The source code and experimental implementation are provided through the anonymous repository accompanying the submission with a detailed reproducible tutorial.

REFERENCES

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pp. 9079–9093, 2024.

Jiaju Chen, Yuxuan Lu, Xiaojie Wang, Huimin Zeng, Jing Huang, Jiri Gesi, Ying Xu, Bingsheng Yao, and Dakuo Wang. Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17389–17416, 2026a.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Tiejin Chen, Huaiyuan Yao, Jia Chen, Evangelos E Papalexakis, and Hua Wei. Every response counts: Quantifying uncertainty of llm-based multi-agent systems through tensor decomposition. *arXiv preprint arXiv:2604.08708*, 2026b.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems*, 38: 101732–101764, 2026.

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.

Sana Ebrahimi, Mohsen Dehghankar, and Abolfazl Asudeh. An adversary-resistant multi-agent llm system via credibility scoring. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1676–1693, 2025.

Yang Feng and Xudong Pan. Sentinelnet: Safeguarding multi-agent collaboration through credit-based dynamic threat detection. In *Proceedings of the ACM Web Conference 2026*, pp. 3135–3146, 2026.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7084–7107, 2025.

Pengfei He, Zhenwei Dai, Xianfeng Tang, Yue Xing, Hui Liu, Jingying Zeng, Qiankun Peng, Shrivats Agrawal, Samarth Varshney, Suhang Wang, et al. Attention knows whom to trust: attention-based trust management for llm multi-agent systems. *arXiv preprint arXiv:2506.02546*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapcoder: Multi-agent code generation for competitive problem solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4912–4944, 2024.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Shiyuan Li, Yixin Liu, Qingsong Wen, Chengqi Zhang, and Shirui Pan. Assemble your crew: Automatic multi-agent communication topology design via autoregressive graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23142–23150, 2026.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Junjun Pan, Yixin Liu, Rui Miao, Kaize Ding, Yu Zheng, Quoc Viet Hung Nguyen, Alan Wee-Chung Liew, and Shirui Pan. Explainable and fine-grained safeguarding of llm multi-agent systems via bi-level graph anomaly detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30491–30506, 2026.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in llm-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12358–12372, 2025.

Akshay Verma, Swapnil Gupta, Deepak Gupta, Prateek Sircar, and Siddharth Pillai. Selene: Selective and evidence-weighted llm debating for efficient and reliable reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 95–104, 2026.

Yihan Xia, Taotao Wang, Shengli Zhang, Zhangyuhua Weng, Bin Cao, and Soung Chang Liew. Hivemind: Contribution-guided online prompt optimization of llm multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29767–29774, 2026.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. *arXiv preprint arXiv:2410.11782*, 2024.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *International Conference on Learning Representations*, volume 2025, pp. 75389–75428, 2025.

Taolin Zhang, Dongyang Li, Chen Chen, Qizhou Chen, Jiuheng Wan, Xiaofeng He, Chengyu Wang, and Richang Hong. Amata: Adaptive multi-agent trajectory alignment for knowledge-intensive question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7879–7900, 2026.

Jun Zhao, Can Zu, Xu Hao, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Longagent: achieving question answering for 128k-token-long documents through multi-agent collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16310–16324, 2024.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33909–33930, 2026.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Language agents as optimizable graphs. *arXiv preprint arXiv:2402.16823*, 2024.

APPENDIX

We provide the appendix, which is organized as following content:

- **Appendix A:** introduces more detailed related works
- **Appendix B:** provides detailed proofs of the theoretical results presented in the main paper, including the analysis of erroneous-majority propagation (B.1), counterfactual topology selection (B.2), and topology identification (B.3).
- **Appendix C:** presents additional experimental results, including additional robustness (C.1) and recovery (C.2) analyses under unreliable multi-agent collaboration.
- **Appendix D:** provides implementation details and source code repository.

A RELATED WORK

A.1 LLM-BASED MULTI-AGENT COLLABORATION

LLM-based multi-agent systems improve complex reasoning by enabling multiple language agents to exchange intermediate solutions and iteratively refine their predictions. Early studies primarily rely on role-based interaction or multi-agent debate, where agents communicate through predefined interaction protocols to aggregate complementary reasoning signals (Li et al., 2023; Du et al., 2024; Hong et al., 2024; Chen et al., 2024). Recent studies further investigate the effectiveness of collaborative reasoning mechanisms such as voting, consensus formation, and structured debate (Choi et al., 2026; Li et al., 2024). These results demonstrate that the interaction protocol and communication structure can substantially influence the effectiveness of multi-agent reasoning. More recently, several approaches formulate multi-agent collaboration as a graph or topology optimization problem. GPTSwarm models language agents as an optimizable graph and learns communication structures for different tasks (Zhuge et al., 2024), while G-Designer dynamically constructs task-adaptive communication topologies (Zhang et al., 2024). Other studies improve communication efficiency by pruning redundant edges or directly generating customized collaboration graphs (Zhang et al., 2025; Shen et al., 2025). While closely related to our work, these methods primarily focus on discovering effective collaboration structures or improving communication efficiency. TrustFlow instead studies information-flow control under ***uncertain agent reliability***, where a communication topology is treated as an intervention whose downstream effect on subsequent collaboration is explicitly evaluated.

A.2 RELIABILITY AND TRUST MODELING IN LLM MULTI-AGENT SYSTEMS

A growing body of work studies the reliability of information exchanged in LLM-based multi-agent systems. One line of research estimates reliability from observable outputs, using signals such as model confidence, self-consistency, response agreement, or group consensus (Chen et al., 2024;

Verma et al., 2026; Zhu et al., 2026; Choi et al., 2026). Such signals provide lightweight indicators of whether an intermediate response should be trusted, but may become less informative when agents exhibit correlated errors or when an incorrect answer dominates the discussion.

Recent studies therefore investigate richer reliability signals at the agent, message, and trajectory levels. A-Trust estimates message- and agent-level trustworthiness from internal attention patterns (He et al., 2025), while MATU quantifies uncertainty in multi-agent reasoning trajectories through tensor decomposition (Chen et al., 2026b). Other approaches use contribution scores, reputation, or anomaly detection to identify unreliable or malicious agents during collaboration (Xia et al., 2026; Feng & Pan, 2026; Pan et al., 2026).

TrustFlow is complementary to these reliability-estimation methods but addresses a different question. Reliability signals indicate ***where*** potential errors may exist, but they do not uniquely determine ***how*** the collaboration process should respond. For the same uncertain agent, preserving its current reasoning, exposing it to a selected peer for verification, or triggering independent rethinking may lead to different downstream outcomes. We therefore model reliability as a latent collaboration state rather than a direct communication decision, and use it to identify candidate interventions whose consequences are subsequently evaluated.

A.3 Adaptive Information-Flow Control

Beyond reliability estimation, recent studies have increasingly recognized communication topology as a key design factor in LLM-based multi-agent systems. Sparse communication structures can maintain competitive reasoning performance while reducing unnecessary interactions (Li et al., 2024), and AgentPrune further removes redundant or potentially harmful communication edges to improve both efficiency and robustness (Zhang et al., 2025). GPTSwarm and G-Designer go beyond manually specified structures by learning or generating task-dependent collaboration graphs (Zhuge et al., 2024; Zhang et al., 2024). More recent work further investigates how communication topology affects the propagation of correct and erroneous information, showing that topology can jointly determine beneficial information diffusion and harmful error amplification (Shen et al., 2025).

TrustFlow shares the general goal of adaptive collaboration, but differs in the criterion used to determine an intervention. Existing topology optimization approaches mainly construct communication structures based on task characteristics, current interaction representations, or learned structural objectives (Zhuge et al., 2024; Zhang et al., 2024; Li et al., 2026). However, under an uncertain collaboration state, multiple plausible interventions may induce different subsequent reasoning trajectories. TrustFlow therefore explicitly evaluates the ***downstream consequences*** of alternative information-flow interventions through counterfactual rollouts. For the same interaction history, candidate topologies are compared according to their topology-conditioned outcomes, and the resulting relative feedback is used to optimize the Controller. In this way, TrustFlow complements topology design with consequence-aware intervention selection, learning not only ***where*** information should flow but also ***which intervention is beneficial*** under the current collaboration state.

B Proof of Theoretical Results

B.1 Proof of Theorem 1

Proof. Let $X_t \in \{0, 1, \dots, N\}$ denote the number of agents supporting the erroneous hypothesis at round t. Under Assumption 1, we model the interaction dynamics as a biased random walk such that, for every non-absorbing state $1 \le k \le N-1$,

$$\Pr(X_{t+1} = k+1 \mid X_t = k) = p, \qquad \Pr(X_{t+1} = k-1 \mid X_t = k) = q,$$

where $p > q$ and $p + q = 1$. The states 0 and N are absorbing and correspond to correct and erroneous consensus, respectively.

Let

$$h_k = \Pr(\tau_N < \tau_0 \mid X_t = k),$$

where τ_0 and τ_N denote the first hitting times of states 0 and N. Then $h_0 = 0$, $h_N = 1$, and for $1 \le k \le N-1$,

$$h_k = p h_{k+1} + q h_{k-1}.$$

Solving this standard difference equation gives

$$h_k = \frac{1-(q/p)^k}{1-(q/p)^N}.$$

Since $p > q$, we have $0 < q/p < 1$ and $1-(q/p)^N < 1$. Hence,

$$h_k \geq 1 - \left(\frac{q}{p}\right)^k.$$

When $k > N/2$,

$$\left(\frac{q}{p}\right)^k < \left(\frac{q}{p}\right)^{N/2},$$

which yields

$$\Pr(\tau_N < \tau_0 \mid X_t = k) > 1 - \left(\frac{q}{p}\right)^{N/2}.$$

Therefore, once the erroneous hypothesis occupies a strict majority, the probability that the system reaches erroneous consensus before correct consensus increases exponentially with the number of agents. □

B.2 Proof of Theorem 2

Proof. For each candidate topology $g \in \mathcal{G}_t$, let

$$R^g_{t,1}, \ldots, R^g_{t,M}$$

denote M independent rollout utilities drawn from the topology-conditioned outcome distribution induced by $P_\Theta(\cdot \mid H_t, g)$. We define the empirical consequence estimate as

$$\widehat{\mu}_t(g) = \frac{1}{M}\sum_{m=1}^{M} R^g_{t,m}.$$

Under Assumption 2, each rollout utility is bounded as

$$R^g_{t,m} \in [R_{\min}, R_{\max}],$$

where

$$B := R_{\max} - R_{\min}$$

denotes the range of the utility function. For any fixed topology $g \in \mathcal{G}_t$, Hoeffding's inequality therefore gives

$$\Pr(|\widehat{\mu}_t(g) - \mu_t(g)| > \epsilon) \leq 2\exp\left(-\frac{2M\epsilon^2}{B^2}\right).$$

Applying a union bound over all $|\mathcal{G}_t|$ candidate topologies yields

$$\Pr\left(\max_{g\in\mathcal{G}_t} |\widehat{\mu}_t(g) - \mu_t(g)| > \epsilon\right) \leq 2|\mathcal{G}_t|\exp\left(-\frac{2M\epsilon^2}{B^2}\right).$$

Setting the right-hand side equal to δ gives

$$\epsilon_M = B\sqrt{\frac{\log(2|\mathcal{G}_t|/\delta)}{2M}}.$$

Hence, with probability at least $1-\delta$, the event

$$\mathcal{E} := \{|\widehat{\mu}_t(g) - \mu_t(g)| \leq \epsilon_M, \quad \forall g \in \mathcal{G}_t\}$$

holds simultaneously for all candidate topologies.

We now condition on the event $\mathcal{E}$. Recall that

$$g_t^\star := \arg\max_{g \in \mathcal{G}_t} \mu_t(g)$$

is the topology with the highest true expected consequence, while

$$\widehat{g}_t := \arg\max_{g \in \mathcal{G}_t} \widehat{\mu}_t(g)$$

is the topology selected by counterfactual evaluation. By the definition of $\widehat{g}_t$,

$$\widehat{\mu}_t(\widehat{g}_t) \geq \widehat{\mu}_t(g_t^\star).$$

Using the uniform concentration event $\mathcal{E}$, we obtain

$$\begin{aligned} \mu_t(\widehat{g}_t) &\geq \widehat{\mu}_t(\widehat{g}_t) - \epsilon_M && (1)\\ &\geq \widehat{\mu}_t(g_t^\star) - \epsilon_M && (2)\\ &\geq \mu_t(g_t^\star) - 2\epsilon_M. && (3) \end{aligned}$$

Therefore, with probability at least $1 - \delta$,

$$\boxed{\mu_t(\widehat{g}_t) \geq \mu_t(g_t^\star) - 2\epsilon_M}.$$

Substituting the definition of ϵ_M gives

$$\mu_t(\widehat{g}_t) \geq \mu_t(g_t^\star) - 2B\sqrt{\frac{\log(2|\mathcal{G}_t|/\delta)}{2M}}.$$

We next compare the counterfactual selector with a topology selected directly from the current reliability belief. Let

$$g_t^{\mathrm{rel}} = \pi_{\mathrm{rel}}(b_t),$$

and define its optimality gap as

$$\Delta_t^{\mathrm{rel}} := \mu_t(g_t^\star) - \mu_t(g_t^{\mathrm{rel}}).$$

Equivalently,

$$\mu_t(g_t^{\mathrm{rel}}) = \mu_t(g_t^\star) - \Delta_t^{\mathrm{rel}}.$$

Combining this identity with the finite-sample lower bound derived above gives

$$\begin{aligned} \mu_t(\widehat{g}_t) - \mu_t(g_t^{\mathrm{rel}}) &\geq \mu_t(g_t^\star) - 2\epsilon_M - \mu_t(g_t^{\mathrm{rel}}) && (4)\\ &= \Delta_t^{\mathrm{rel}} - 2\epsilon_M. && (5) \end{aligned}$$

Consequently, whenever

$$\Delta_t^{\mathrm{rel}} > 2\epsilon_M,$$

we have

$$\mu_t(\widehat{g}_t) > \mu_t(g_t^{\mathrm{rel}})$$

on the event $\mathcal{E}$. Since $\Pr(\mathcal{E}) \geq 1 - \delta$, it follows that

$$\Pr\big(\mu_t(\widehat{g}_t) > \mu_t(g_t^{\mathrm{rel}})\big) \geq 1 - \delta.$$

Thus, provided that the optimality gap of the reliability-only selection exceeds the finite-sample estimation error, counterfactual evaluation selects a topology with strictly higher expected consequence with probability at least $1 - \delta$. □

B.3 Proof of Theorem 3

Proof. For any suboptimal topology $g \neq g_t^\star$, define $\Delta_g := \mu_t(g_t^\star) - \mu_t(g)$. By definition, $\Delta_g \geq \Delta_t$.

The counterfactual selector fails to identify $g_t^\star$ only if there exists some suboptimal topology g such that $\widehat{\mu}_t(g) \geq \widehat{\mu}_t(g_t^\star)$. For this event to occur, at least one of the following deviations must hold:

$$\widehat{\mu}_t(g) - \mu_t(g) \geq \frac{\Delta_g}{2}, \qquad \text{or} \qquad \mu_t(g_t^\star) - \widehat{\mu}_t(g_t^\star) \geq \frac{\Delta_g}{2}.$$

Since each rollout utility lies in an interval of length B, Hoeffding's inequality gives

$$\Pr\left(\widehat{\mu}_t(g) - \mu_t(g) \geq \frac{\Delta_g}{2}\right) \leq \exp\left(-\frac{M\Delta_g^2}{2B^2}\right),$$

and the same bound holds for the deviation of $\widehat{\mu}_t(g_t^\star)$. Therefore,

$$\Pr(\widehat{\mu}_t(g) \geq \widehat{\mu}_t(g_t^\star)) \leq 2\exp\left(-\frac{M\Delta_g^2}{2B^2}\right) \leq 2\exp\left(-\frac{M\Delta_t^2}{2B^2}\right).$$

Applying a union bound over all $|\mathcal{G}_t| - 1$ suboptimal topologies yields

$$\Pr(\widehat{g}_t \neq g_t^\star) \leq 2(|\mathcal{G}_t| - 1)\exp\left(-\frac{M\Delta_t^2}{2B^2}\right).$$

Hence, requiring

$$M \geq \frac{2B^2}{\Delta_t^2}\log\frac{2(|\mathcal{G}_t| - 1)}{\delta}$$

ensures $\Pr(\widehat{g}_t \neq g_t^\star) \leq \delta$, or equivalently, $\Pr(\widehat{g}_t = g_t^\star) \geq 1 - \delta$.

Finally, if $g_t^{\text{rel}} \neq g_t^\star$, then by uniqueness of $g_t^\star$, $\mu_t(g_t^\star) > \mu_t(g_t^{\text{rel}})$. Therefore, whenever $\widehat{g}_t = g_t^\star$, we have $\mu_t(\widehat{g}_t) > \mu_t(g_t^{\text{rel}})$. The latter event thus holds with probability at least $1 - \delta$. □

C Additional Experiments

C.1 Additional Robustness Analysis

We further evaluate the robustness of TrustFlow when unreliable information is introduced into multi-agent collaboration. Specifically, we consider two challenging settings in which either one or two agents provide incorrect responses at the beginning of the collaboration process. A robust collaboration strategy should prevent these erroneous responses from being amplified through subsequent interactions and preserve useful information from reliable agents. As shown in Fig. 9, we compare different multi-agent reasoning methods under different numbers of initially incorrect agents on TruthfulQA.

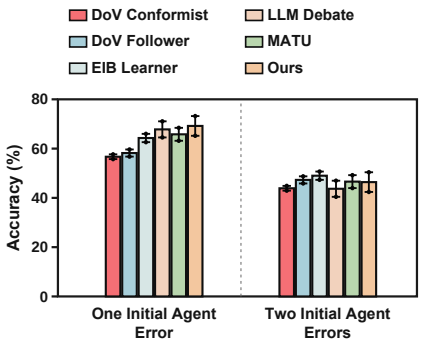
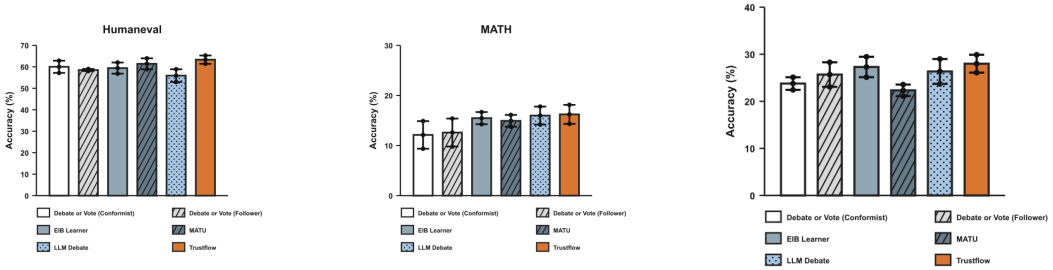


Figure 9: Robustness analysis on TruthfulQA under different numbers of initially incorrect agents.

Analysis. Increasing the number of initially incorrect agents generally leads to performance degradation across different methods, highlighting the difficulty of maintaining reliable collaboration when erroneous information becomes dominant. TrustFlow maintains strong performance when only one agent is initially incorrect, suggesting that adaptive information-flow control can effectively restrict the propagation of isolated erroneous responses while preserving useful information from reliable agents. When two agents are initially incorrect, the performance differences among methods become smaller, as reliable information is substantially more limited at the beginning of collaboration. Overall, these results indicate that TrustFlow improves robustness by regulating the propagation of unreliable information rather than simply increasing communication among agents.



(a) Recovery analysis on HumanEval. **(b)** Recovery analysis on MATH. **(c)** Recovery analysis on TruthfulQA.

Figure 10: Additional recovery analysis across three reasoning benchmarks. We examine how different collaboration strategies recover from unreliable intermediate states and prevent erroneous information from dominating subsequent reasoning.

C.2 Additional Recovery Analysis

We further analyze whether different collaboration strategies can recover from erroneous intermediate states during multi-agent reasoning. In particular, we focus on the ability of the system to transform initially unreliable collaboration states into correct final outcomes while avoiding harmful propagation that may mislead otherwise correct agents.

Analysis. Fig. 10 further demonstrates the effectiveness of adaptive information-flow control in recovering from unreliable collaboration states. Across HumanEval, MATH, and TruthfulQA, TrustFlow consistently achieves strong recovery performance compared with alternative collaboration strategies. The improvement is particularly evident on HumanEval and TruthfulQA, where inappropriate information exchange can easily propagate an incorrect intermediate solution to other agents. By conditioning its interaction decisions on the inferred reliability state, TrustFlow can selectively restrict potentially harmful information while preserving communication that contributes to correction.

More importantly, these results complement the robustness analysis in Fig. 10. While the robustness experiments evaluate whether the system can resist initially introduced errors, the recovery experiments examine whether the collaboration process can actively escape from erroneous states once they have occurred. Together, the two analyses show that TrustFlow not only mitigates error propagation but also facilitates the recovery of collaborative reasoning when intermediate agent responses become unreliable.

D Implementation Details

Training Settings This section provides additional implementation details. We adopt a two-stage training procedure. We first train the belief estimator using interaction trajectories generated by frozen language agents. The estimator is then frozen, together with the underlying answering agents, and only the LoRA parameters of the topology controller are optimized using GRPO.The datasets used for controller training are disjoint from those used to train the belief estimator.All experiments were conducted on NVIDIA A100 GPUs.

Candidate Topology Sampling for Larger Agent Groups. To control the computational cost of counterfactual evaluation as the number of agents increases, we adopt a bounded candidate sampling strategy for larger agent groups. Specifically, when $N > 6$, we select the three agents with the highest estimated unreliability scores ($k = 3$) as intervention anchors. For each selected agent, we sample 10 candidate information-flow topologies involving different Keep, Verify, and Rethink configurations. Consequently, at most 30 candidate topologies are evaluated for each interaction state, preventing the counterfactual evaluation cost from growing combinatorially with N.

Code link:https://anonymous.4open.science/r/TrustFlow-EB33