Conformity Dynamics in LLM Multi-Agent Systems: The Roles of Topology and Self–Social Weighting

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly instantiated as interacting agents in multi-agent systems (MAS), where collective decisions emerge through social interaction rather than independent reasoning. A fundamental yet underexplored mechanism in this process is conformity, the tendency of agents to align their judgments with prevailing group opinions. This paper presents a systematic study of how network topology shapes conformity dynamics in LLM-based MAS through a misinformation detection task. We introduce a confidence-normalized pooling rule that controls the trade-off between self-reliance and social influence, enabling comparisons between two canonical decision paradigms: Centralized Aggregation and Distributed Consensus. Experimental results demonstrate that network topology critically governs both the efficiency and robustness of collective judgments. Centralized structures enable immediate decisions but are sensitive to hub competence and exhibit same-model alignment biases. In contrast, distributed structures promote more robust consensus, while increased network connectivity speeds up convergence but also heightens the risk of wrong-but-sure cascades, in which agents converge on incorrect decisions with high confidence. These findings characterize the conformity dynamics in LLM-based MAS, clarifying how network topology and self–social weighting jointly shape the efficiency, robustness, and failure modes of collective decision-making. The code and dataset are available at: `Topology-of-Multi-Agent-Systems`.

1 Introduction

Large Language Models (LLMs) have been increasingly instantiated as interacting agents within Multi-Agent Systems (MAS) (Wei et al., 2025). Across a wide range of applications, including collaborative problem solving (Du et al., 2024), complex reasoning (Zhang and Xiong, 2025) and misinformation detection (Li et al., 2025), the effectiveness of collective decision-making depends not only on the competence of individual agents, but also on the social dynamics induced by their interactions (Ghoshal et al., 2025; Cisneros-Velarde, 2025). A central mechanism underlying these dynamics is conformity, the tendency of agents to adjust their judgments toward majority opinions. Prior studies suggest that conformity can reduce idiosyncratic noise and facilitate coordination (Choi et al., 2025). However, excessive conformity may also cause information cascades, leading groups to converge on incorrect conclusions with high confidence (Bikhchandani et al., 2024; Pinheiro and Vasconcelos, 2025).

Existing research on MAS has primarily emphasized task efficacy, focusing on protocol design, role specialization, and coordination efficiency (Agashe et al., 2025; Grötschla et al., 2025). For example, multi-agent debate frameworks can shift individual judgments toward majority positions, improving reliability while reinforcing systematic biases (Han et al., 2025). Related work on persuasion further suggests that LLM agents tend to imitate dominant argumentative patterns during the interaction (Argyle et al., 2025). While conformity has been studied in computational opinion dynamics (Helfmann et al., 2023; Ding et al., 2025; Han and Tang, 2025), existing approaches still lack an explicit treatment of how agents generate judgments during decision processes, and in particular how interaction topology and neighbor effect jointly govern the propagation, aggregation, and amplification of agent confidence.

In this study, we bridge this gap by investigating how network topology modulates conformity through a binary misinformation detection task. We propose a confidence-normalized update rule governed by a global self-weighting parameter α, which balances an agent's self-reliance against their

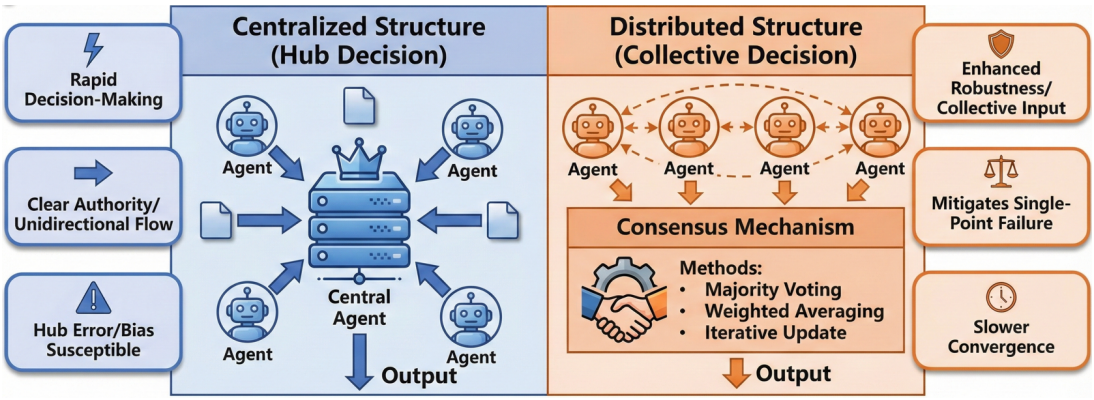

Figure 1: Illustration of centralized (hub-based) and distributed (collective) multi-agent structures, highlighting their decision-making mechanisms, relative advantages, and inherent limitations in collective inference.

neighborhood influence. Specifically, we compare two distinct decision modes: (i) **Centralized Aggregation** (e.g., star networks), where a hub synthesizes a collective decision in a single round; and (ii) **Distributed Consensus** (e.g., sparse rings to complete graph), where influence propagates iteratively. This allows us to disentangle one-step hub dominance from multi-round interaction. Our contributions are summarized as follows:

(1) Topology as a determinant of conformity. We show that network structure systematically shapes conformity by trading off convergence speed and robustness. Centralized structures concentrate influence, yielding immediate decisions but making outcomes highly dependent on hub competence and model homogeneity; Distributed structures diffuse influence and support more robust consensus formation, while increased connectivity accelerates convergence and may facilitate high-confidence error cascades.

(2) A transparent confidence-normalized pooling rule. We propose a transparent and generalizable update mechanism that introduces a global parameter to explicitly regulate social influence. By integrating agent-level confidence into the pooling process, the rule yields bounded belief scores, stable binarization, and well-controlled conformity dynamics, providing a principled basis for different decision paradigms.

(3) Empirical insights on robustness and risk. Experiments on current fact-checking datasets reveal that system reliability is heavily contingent on hub competence and majority composition. We further characterize the dual nature of conformity, showing that it enhances accuracy under reliable conditions but can induce confident errors when early majority signals are incorrect.

2 Related Works

2.1 MAS and Collective Decision-Making

Early research on MAS primarily focused on distributed optimization and consensus formation under idealized assumptions, where agents were modeled as homogeneous entities with limited reasoning capacity (Bao et al., 2022; Amirkhani and Barshooi, 2022). The recent integration of LLMs expands this paradigm by endowing agents with advanced reasoning, enabling MAS to address complex, unstructured decision-making problems such as multi-step problem solving (Chen et al., 2024) and automated fact-checking (Han et al., 2025). Despite these advances, most existing studies emphasize aggregate performance metrics, leaving the role of social dynamics such as conformity in shaping collective reliability largely unexamined.

2.2 Conformity and Opinion Dynamics

Conformity, defined as the tendency of individuals to align their judgments with a perceived majority, has been widely studied in social psychology (Capuano and Chekroun, 2024). In computational settings, opinion dynamics models such as DeGroot averaging (Dong et al., 2024) and bounded-confidence mechanisms (Li and Porter, 2023) formalize how local interactions give rise to collective consensus and have been used to explain phenomena including information cascades and polarization (Shirzadi et al., 2025). However, these models operate at an abstract level and do not capture the semantic reasoning or contextual understanding characteristic of modern LLM agents (Aouini and Loubani, 2025). Consequently, how classical conformity theories extend to LLM-based MAS, where judgments are produced through generative processes, remains insufficiently explored.

2.3 Network Topology

Network topology is a fundamental determinant of how influence propagates and consensus forms in multi-agent systems (Cheng et al., 2021; Amirkhani and Barshooi, 2022). **Centralized aggregation** structures, such as star or hierarchical graphs, concentrate influence and enable rapid decision-making, but render collective outcomes highly sensitive to the reliability of central agents. By contrast, **Distributed consensus** structures, including ring and complete graphs, diffuse influence across agents, enhancing robustness to local noise at the expense of slower convergence. Latest studies have further examined how connectivity modulates communication cost, convergence speed, and robustness (Da et al., 2025; Wang et al., 2025; Yang et al., 2025). Despite these efforts, existing work has yet to clarify the interplay between interaction topology, agent confidence, and conformity in shaping collective outcomes in LLM-driven MAS.

3 Methodology

3.1 Agent Decision Model

Building on classical opinion dynamics and weighted consensus frameworks (Anderson and Ye, 2019; Li and Porter, 2023; Dong et al., 2024), we formalize misinformation detection in MAS as a binary collective decision problem. At each interaction round t, agent i produces a binary judgment $y_i^{(t)} \in 0, 1$ (with 0 denoting True and 1 denoting False), accompanied by a confidence score $p_i^{(t)}$ that quantifies the agent's self-assessed reliability.

Update rule. Agents update an internal support score using confidence-normalized pooling:

$$s_i^{(t+1)} = \frac{\alpha\, p_i^{(t)} y_i^{(t)} + (1-\alpha) \sum_{j \in N_i} p_j^{(t)} y_j^{(t)}}{\alpha\, p_i^{(t)} + (1-\alpha) \sum_{j \in N_i} p_j^{(t)} + \varepsilon}, \quad (1)$$

where N_i denotes the neighbor set of agent i, and ε ensures numerical stability. The dynamics are governed by two parameters: $\alpha \in [0, 1]$ is a fixed hyperparameter that balances self-reliance and peer influence, and $p_i^{(t)} \in [0, 1]$ is a self-reported confidence score generated by LLMs at each round, modulating both the persistence of its judgment and the influence on neighbors. By construction, $s_i^{(t+1)} \in (0, 1)$, while in high self-reliance (large α), strong confident neighbor signals can still sway the decision.

Binary readout. The score is mapped to a binary label using a fixed threshold τ:

$$y_i^{(t+1)} = \mathbb{1}\left[s_i^{(t+1)} \geq \tau\right], \quad (2)$$

where $\tau = 0.5$ by default, so $s_i^{(t+1)} < 0.5$ yields True (0) and $s_i^{(t+1)} \geq 0.5$ yields False (1), ensuring class symmetry and stable, interpretable binarization across tasks and confidence distributions.

3.2 Prompt Design

Each agent receives a structured prompt comprising three core components: (i) a concise background profile, automatically generated by LLMs, situating the claim within its domain context; (ii) a task description instructing the agent to evaluate the claim based on the provided profile and its own reasoning; and **(iii) output requirements specifying a binary label $y_i^{(t)} \in \{0, 1\}$, a confidence score $p_i \in [0, 1]$, and a brief justification**. These three outputs are then incorporated into the update rule in Eq. (1) to revise the agent's belief and guide subsequent decisions. Complete prompts and pseudocode are provided in Appendix A.

3.3 Network Topologies

We analyze conformity dynamics across two distinct topological paradigms: **Centralized Aggregation** and **Distributed Consensus**. Centralized structures rely on immediate, hub-mediated consolidation, and distributed architectures foster iterative, emergent consensus. To ensure comparability, all topologies are instantiated with seven fixed agents ($N = 7$). Detailed explanations of the evaluation metrics are provided in Appendix D.

3.3.1 Centralized Aggregation

Centralized Aggregation is characterized by a unidirectional upward information flow from peripheral agents to a central authority, whose synthesis determines the collective outcome. We examine two representative configurations (Figure 2):

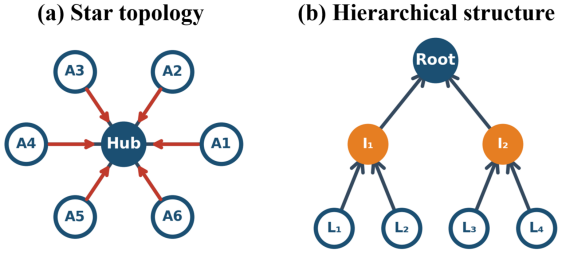

Figure 2: Centralized Aggregation Topology.

Topology	α	GPT-3.5			GPT-4o			Llama3.3		
		CA	PA	CPC	CA	PA	CPC	CA	PA	CPC
Star	0.00	0.67	0.62	0.62	0.69	0.65	0.63	0.66	0.62	0.65
	0.25	0.73	0.68	0.73	0.75	0.71	0.77	0.74	0.72	0.68
	0.50	0.76	**0.71**	0.75	0.77	0.73	0.82	0.76	0.74	0.84
	0.75	**0.80**	0.70	**0.79**	**0.82**	**0.77**	**0.84**	**0.80**	**0.80**	**0.87**
	1.00	0.75	0.69	0.74	0.77	0.75	0.75	0.78	0.74	0.73
	No-weight	0.69	0.64	0.58	0.71	0.66	0.60	0.70	0.65	0.59
Hierarchical	0.00	0.66	0.63	0.67	0.68	0.67	0.67	0.70	0.65	0.61
	0.25	0.70	0.67	0.70	0.72	0.69	0.71	0.73	0.70	0.64
	0.50	0.72	0.68	**0.77**	0.76	0.70	0.80	0.77	0.71	**0.82**
	0.75	**0.75**	**0.71**	**0.77**	**0.78**	0.73	**0.82**	**0.79**	**0.75**	0.80
	1.00	0.72	0.69	0.74	0.75	**0.74**	0.78	0.75	0.72	0.79
	No-weight	0.68	0.65	0.60	0.70	0.66	0.62	0.71	0.67	0.63

Table 1: Centralized Aggregation under the single-round protocol. Results are reported in terms of central accuracy (CA), peripheral accuracy (PA), and center–periphery consistency (CPC). Best and second-best results within each topology are highlighted in **bold** and underline, respectively (ties are marked). The *No-weight* setting removes confidence-weighted aggregation and the global self-weighting parameter. Across models and topologies, $\alpha = 0.75$ most often achieves the best performance.

(a) Star Network: Six peripheral nodes transmit judgments directly to a central hub without lateral interaction.

(b) Hierarchical Structure: A three-layered tiered architecture where inputs from leaf nodes are aggregated by intermediate agents before reaching the root.

Protocol. In both structures, decision-making is executed in a single update round. Peripheral or lower-tier agents submit judgments to the hub or root, which integrates these inputs to generate a final output. This central output is adopted as the group's collective decision.

Metrics. We assess system reliability and hub influence with three metrics: **(1) Central Accuracy (CA):** Correctness of the hub/root node's final judgment against ground truth. **(2) Peripheral Accuracy (PA):** Mean correctness of all non-central nodes. **(3) Center–Periphery Consistency (CPC):** The proportion of peripheral nodes aligned with the central decision, quantifying immediate conformity intensity.

3.3.2 Distributed Consensus

Distributed Consensus distributes influence symmetrically, preventing any single node from dominating the decision process. We explore a connectivity spectrum ranging from sparse rings to complete graphs (Figure 3):

(a) Sparse ring (2 neighbors): Agents interact strictly with their immediate predecessor and successor.

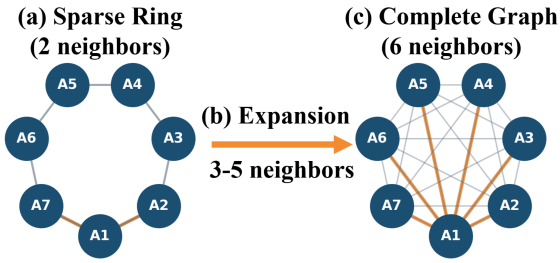

Figure 3: Distributed Consensus Topology.

(b) Expanded rings (3–5 neighbors): Agents connect to a broader local vicinity, accelerating information diffusion.

(c) Complete graph (6 neighbors): Agents connect to all others, maximizing potential conformity intensity.

Protocol. Decision-making operates via iterative rounds. In each step, agents exchange states with neighbors and update their internal confidence and judgment according to Eq. (1). The process terminates upon full consensus (all agents hold the same judgment) or when the maximum round limit $T_{\max}$ is reached.

Metrics. We evaluate the dynamics of emergent consensus with four metrics: **(1) Final Accuracy (FA)**: Concordance of the group decision (unanimous or majority view at $T_{\max}$) with the ground truth. **(2) Time-to-Consensus (TTC)**: The number of rounds required to reach unanimity. It is unde-

Neighbors	α	GPT-3.5				GPT-4o				Llama3.3			
		FA	TTC	ACI	TT	FA	TTC	ACI	TT	FA	TTC	ACI	TT
2	0.00	0.68	7.0	0.58	9.5k	0.70	6.4	0.62	9.2k	0.67	6.2	0.60	7.9k
	0.25	0.70	6.5	0.61	8.9k	0.73	6.2	0.67	9.0k	0.69	5.9	0.66	7.5k
	0.50	0.72	6.3	0.71	8.7k	0.75	5.7	0.72	8.3k	0.73	5.8	0.71	7.4k
	0.75	**0.75**	5.8	**0.75**	8.2k	**0.79**	5.1	**0.77**	7.6k	**0.78**	5.6	**0.78**	7.2k
	1.00	0.74	**5.3**	0.68	**7.5k**	0.78	**5.0**	0.75	**7.5k**	0.76	**5.1**	0.73	**6.5k**
	No-weight	0.69	7.4	0.55	9.8k	0.71	6.9	0.57	9.9k	0.70	6.7	0.56	8.3k
3	0.00	0.70	7.2	0.60	10.7k	0.70	6.5	0.63	10.2k	0.68	6.3	0.61	8.8k
	0.25	0.72	6.5	0.64	9.8k	0.74	5.8	0.70	9.3k	0.70	5.9	0.69	8.0k
	0.50	0.75	5.8	0.73	9.0k	0.75	5.3	0.77	8.7k	0.76	5.6	0.74	7.8k
	0.75	**0.78**	5.1	**0.77**	8.2k	**0.82**	4.7	**0.81**	7.8k	**0.79**	5.0	**0.80**	7.2k
	1.00	0.77	**4.9**	0.74	**8.0k**	0.80	**4.5**	0.79	**7.7k**	0.78	**4.8**	0.77	**7.1k**
	No-weight	0.71	6.8	0.58	10.2k	0.73	6.2	0.60	9.6k	0.72	6.1	0.59	8.4k
4	0.00	0.69	6.9	0.61	11.2k	0.71	6.3	0.63	11.0k	0.70	6.0	0.62	9.1k
	0.25	0.71	6.1	0.69	10.1k	0.74	5.4	0.73	9.6k	0.72	5.7	0.73	8.4k
	0.50	0.74	5.0	0.74	8.8k	0.81	4.7	0.77	8.6k	0.79	5.0	0.76	7.5k
	0.75	**0.78**	4.3	**0.79**	7.9k	**0.83**	4.1	**0.81**	7.9k	**0.81**	4.2	**0.82**	6.9k
	1.00	0.77	**4.0**	0.75	**7.4k**	0.82	**3.9**	0.80	**7.6k**	0.80	**4.1**	0.79	**6.8k**
	No-weight	0.72	6.2	0.60	10.2k	0.74	5.7	0.62	9.9k	0.73	5.6	0.61	8.4k
5	0.00	0.69	6.8	0.59	11.7k	0.71	6.2	0.62	11.4k	0.68	6.1	0.60	9.8k
	0.25	0.72	5.5	0.70	10.0k	0.78	5.3	0.73	10.3k	0.74	5.1	0.74	7.9k
	0.50	0.74	4.5	0.76	8.6k	0.81	4.1	0.79	8.8k	0.82	4.4	0.78	7.6k
	0.75	**0.75**	3.7	**0.81**	7.6k	**0.84**	3.5	**0.83**	7.8k	**0.84**	3.4	**0.84**	6.6k
	1.00	0.74	**3.2**	0.78	**6.9k**	0.83	**3.0**	0.81	**7.0k**	0.81	**3.1**	0.80	**6.2k**
	No-weight	0.73	5.8	0.63	10.4k	0.75	5.2	0.65	10.3k	0.74	5.3	0.64	8.5k
6	0.00	0.72	6.3	0.62	11.4k	0.74	5.8	0.65	11.1k	0.71	5.7	0.63	9.0k
	0.25	0.75	5.0	0.74	9.4k	0.77	4.6	0.77	9.3k	0.78	4.7	0.78	7.3k
	0.50	0.75	4.0	0.80	7.8k	0.82	3.8	0.77	8.0k	0.81	4.2	0.81	6.7k
	0.75	**0.78**	3.8	**0.81**	7.7k	**0.85**	3.4	**0.85**	7.2k	**0.83**	3.2	**0.82**	6.1k
	1.00	0.77	**3.2**	0.78	**6.7k**	0.80	**2.9**	0.83	**6.8k**	0.82	**3.0**	0.81	**5.6k**
	No-weight	0.74	5.4	0.66	10.1k	0.76	5.0	0.68	9.8k	0.75	5.1	0.67	7.8k

Table 2: Results of Distributed Consensus under varying neighbor counts m and self-weighting. Performance is reported in terms of final accuracy (FA), time-to-consensus (TTC), average conformity index (ACI), and total tokens (TT). For each neighbor-count block, the best and second-best results are highlighted in **bold** and underline, respectively (ties are marked). The *No-weight* setting removes confidence-weighted aggregation and the global self-weighting parameter. Across models and topologies, $\alpha = 0.75$ most often achieves the best performance.

fined if the system fails to converge within $T_{\max}$. **(3) Conformity Index (CI)**: Degree of within-group agreement at round t, defined as the proportion of agents adopting the majority label. Since judgments are binary, CI is bounded in $[0.5, 1]$, where CI $= 1$ indicates unanimity and CI $= 0.5$ corresponds to an evenly split group. We report **Average CI (ACI)** as the mean CI over rounds up to $T_{\max}$. **(4) Total Tokens (TT)**: The cumulative cost, summing token usage across all agent outputs and interaction rounds.

4 Experiment

4.1 Experimental Setup

Dataset. We collect Snopes25, a new benchmark comprising 448 real-world claims (252 false, 196 true) fact-checked by Snopes editors. All claims are from January to June 2025 to minimize potential data contamination from pre-trained knowledge.

Implementation. Experiments are conducted using two proprietary models, GPT-3.5 (OpenAI, 2023) and GPT-4o (OpenAI, 2024), and one open-source model, Llama3.3-70B-Instruct (AI@Meta, 2024) to ensure all models' training cutoff precede our evaluation data. Detailed hyperparameters and zero-shot baselines are provided in Appendix B. For **Centralized Aggregation**, decision-making is restricted to a single update round, capturing immediate conformity to the hub or root node. For **Distributed Consensus**, agents iteratively exchange judgments for up to $T_{\max} = 10$ rounds or until consensus is reached.

Parameter Configurations. We consider five values of α: 0 (fully conformist), 0.25 (socially influenced), 0.50 (balanced), 0.75 (self-reliant), and 1 (fully independent), capturing different predispositions toward peer influence versus self-reliance. All reported metrics represent the mean of 10 in-

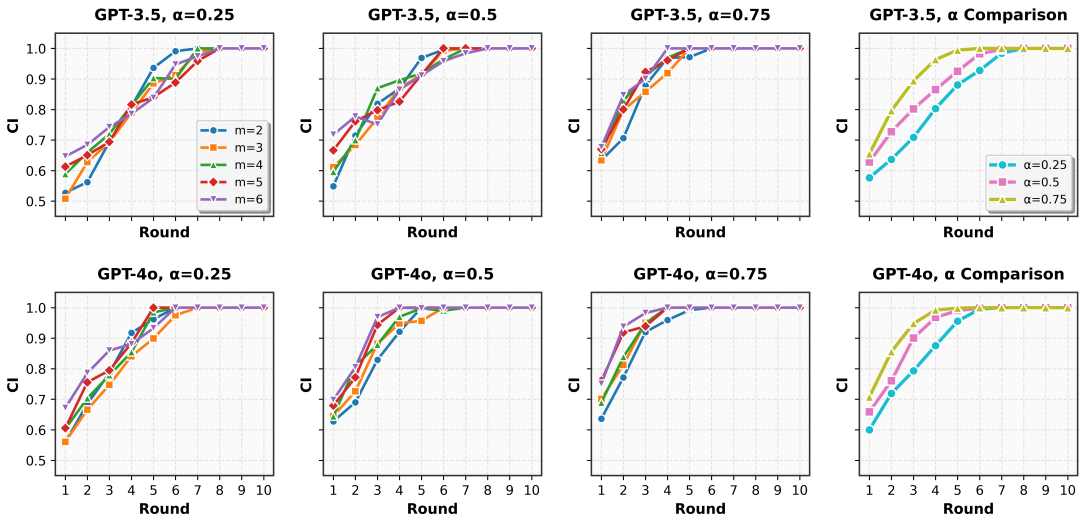

Figure 4: Temporal evolution of the Conformity Index across iterative rounds under varying network densities and self–social weighting, revealing rapid early alignment and diminishing marginal gains in dense structures.

dependent runs, with standard errors consistent at < 0.09. To isolate the contribution of our confidence-weighting mechanism, we introduce a *No-weight* baseline that removes both the self-weighting parameter α and the confidence score p, forcing agent updates to depend solely on the textual context of neighbor interactions and isolating the effect of explicit confidence weighting.

4.2 Results on Centralized Aggregation

We first evaluate **Centralized Aggregation**, and the results are summarized in Table 1. Across both topologies and all models, central accuracy (CA) increases monotonically with the self-weighting parameter α, indicating that stronger self-reliance enables the hub to better filter peripheral noise. For example, in the star topology, CA for GPT-3.5 improves from 0.67 at $\alpha = 0$ to 0.80 at $\alpha = 0.75$. Hierarchical structures follow the same trend, though with attenuated improvements due to information dilution across levels.

In contrast, peripheral accuracy (PA) remains largely stable across α, indicating that variations in self–social weighting primarily influence the quality of the hub's aggregation rather than improving local agent correctness. Center–periphery consistency (CPC) increases with α across models and topologies, suggesting that as the hub places greater emphasis on its own confidence-weighted judgment, its final decision becomes more stable and, given sufficient hub competence, more likely to align with peripheral votes.

4.3 Results on Distributed Consensus

We next turn to **Distributed Consensus**, and Table 2 summarizes performance across varying network densities and self-social weighting. Final Accuracy (FA) generally improves with connectivity m and often peaks at moderate-to-high self-reliance ($\alpha = 0.75$). For instance, with GPT-3.5 under a socially influenced setting ($\alpha = 0.25$), increasing connectivity from sparse rings ($m = 2$) to complete graphs ($m = 6$) raises accuracy from 0.70 to 0.75. Denser networks facilitate rapid information propagation and thus accelerate convergence, whereas decreasing α tends to strengthen conformity pressure and can further speed up consensus, albeit at the risk of amplifying early biases. In sparse networks ($m = 2$) with strong peer influence ($\alpha = 0.25$), conformity remains moderate, whereas in highly connected regimes ($m = 6$), groups exhibit high homogeneity.

Without the confidence-weighted mechanism, FA degrades uniformly across all topologies, and simply increasing connectivity fails to recover the gains of the full model. While increasing neighbors inflates the volume of messages per round, higher m substantially reduces the number of rounds required to reach consensus. In summary, sparse networks preserve diversity and mitigate premature convergence, while complete graphs accelerate consensus but may increase susceptibility to information cascades when early signals are biased.

Figure 4 further illustrates the temporal evolution of the Conformity Index (CI) across rounds under different neighbor counts m and self–social

Setting	Model Assignment			Performance				
	Hub	Left	Right	CA	PA_L	PA_R	CPC_L	CPC_R
Cap-H1	GPT-4o	GPT-4o	GPT-3.5	0.82	0.79	0.67	0.87	0.74
Cap-H2	GPT-3.5	GPT-3.5	GPT-4o	0.77	0.73	0.77	0.81	0.69
Cap-H3	Llama3.3-70B	Llama3.3-70B	Llama3.3-8B	0.80	0.76	0.70	0.86	0.78
Cap-H4	Llama3.3-8B	Llama3.3-8B	Llama3.3-70B	0.76	0.71	0.75	0.80	0.72
Type-H1	GPT-4o	GPT-4o	Llama3.3-70B	0.82	0.75	0.73	0.88	0.76
Type-H2	Llama3.3-70B	Llama3.3-70B	GPT-4o	0.79	0.72	0.76	0.84	0.77

Table 3: Heterogeneous Centralized Aggregation. The Left branch is aligned with the hub (same backbone), while the Right branch is assigned a different model. We report Central Accuracy (CA), branch-level Peripheral Accuracy (PA_L, PA_R), and Center–Periphery Consistency (CPC_L, CPC_R). Centralized performance is primarily driven by hub competence rather than peripheral model strength.

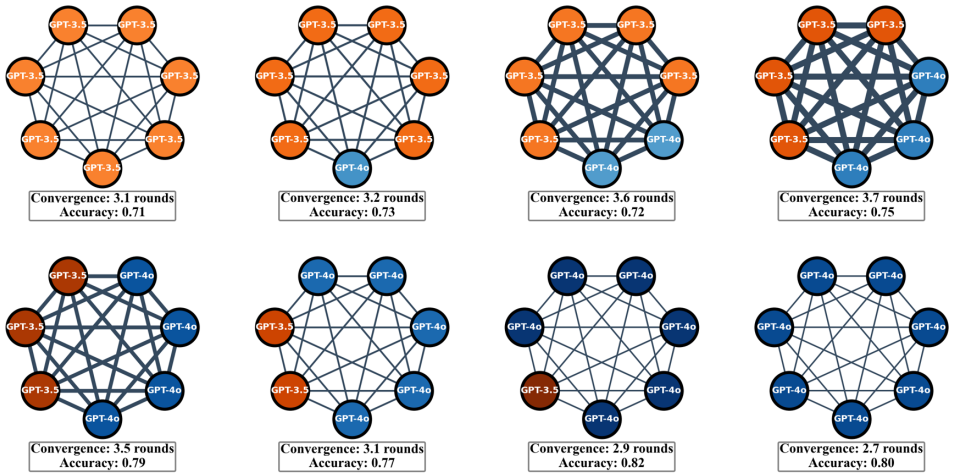

Figure 5: Distributed Consensus with heterogeneous model composition. Darker node colors indicate higher accuracy, while thicker edges signify prolonged time-to-consensus. Both the proportion and structure of stronger agents jointly influence convergence speed and final collective accuracy.

weighting. CI rises markedly faster in denser networks, indicating that quicker consolidation of group alignment once a dominant opinion emerges. Specifically, sparse rings ($m = 2$) typically start from 0.55-0.65 and may take six or more rounds to approach unanimity, whereas denser structures ($m \geq 4$) exceed 0.75 by the second round and nearly converge by the fourth. Across all settings, the steepest CI increase occurs within the first four rounds, followed by a slower consolidation phase until consensus, and the gains from increasing α or enlarging m exhibit diminishing returns once networks are sufficiently dense.

5 Discussion: Factors Influencing the Conformity in MAS

5.1 Model Heterogeneity

To characterize model heterogeneity in LLM-based MAS, agents are instantiated with different LLM backbones. In **Centralized Aggregation**, the hub shares its backbone with the Left branch (the *Aligned Branch*), while the Right branch is instantiated with a different model (the *Opposing Branch*). We fix the self–social weighting at $\alpha = 0.75$ and report hub-level Central Accuracy (CA), branch-level Peripheral Accuracy (PA_L, PA_R), and Center–Periphery Consistency (CPC_L, CPC_R) over 10 independent runs per claim.

Table 3 reveals two salient phenomena associated with centralized conformity. First, a consistent *same-model alignment bias* emerges: the hub aligns more frequently with the Aligned Branch than with the Opposing Branch, evidenced by CPC_L being consistently higher than CPC_R across all settings. This pattern suggests that the hub is more receptive to peer rationales that resemble its own reasoning style and inductive biases (Laurito et al., 2025). Second, system reliability is predominantly capped by the hub's competence: CA is higher when the hub is instantiated with a stronger model, even when the periphery contains a stronger

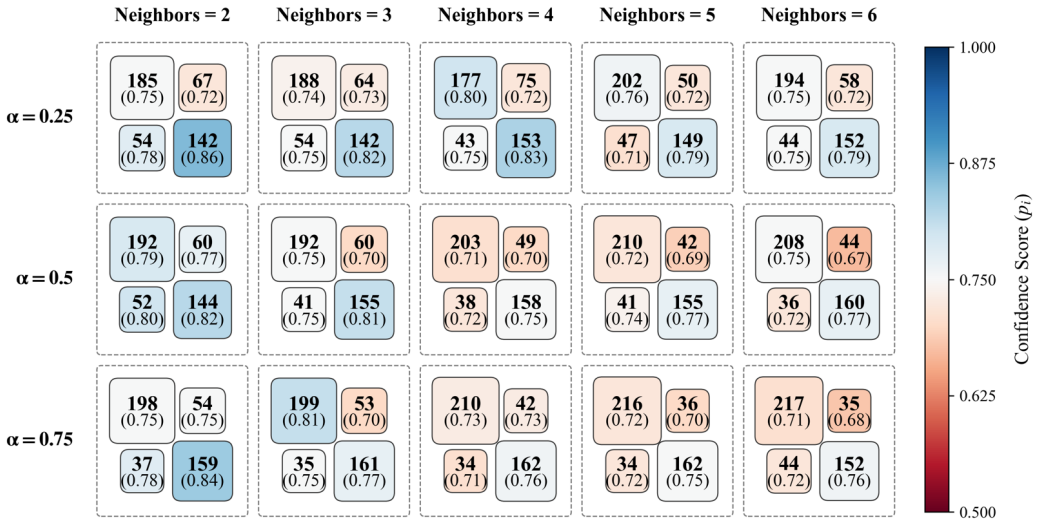

Figure 6: Confusion-matrix heatmaps of GPT-4o under **Distributed Consensus**, varying neighbor counts and α values. Each cell reports group-level outcomes: the number of correctly classified true claims (top-left), correctly classified false claims (top-right), misclassified true claims (bottom-left), and misclassified false claims (bottom-right). Cell size reflects the frequency, while color encodes the mean confidence score p_i. Higher connectivity and lower self-reliance produce more high-confidence misclassifications, highlighting a key failure mode of conformity.

model on the Opposing Branch. Together, these results highlight a vulnerability of centralized MAS: peripheral diversity does not reliably compensate for central incompetence; instead, the hub can act as a selective filter that amplifies aligned inputs while down-weighting divergent perspectives.

We further examine heterogeneity in **Distributed Consensus** by varying the ratio of GPT-3.5 to GPT-4o (0:7 to 7:0) in complete graphs. Experiments are conducted on 50 claims with 10 runs per claim at $\alpha = 0.75$. As illustrated in Figure 5, heterogeneous groups converge more slowly than homogeneous ones, consistent with increased deliberative friction induced by divergent reasoning priors and confidence calibration. In contrast, homogeneous groups converge more rapidly, and groups in which GPT-4o agents constitute the majority consistently achieve higher final accuracy. This indicates that while homogeneity primarily governs convergence speed, the accuracy of the emergent consensus is ultimately determined by the capability of the prevailing model class.

5.2 Double-Edged Effect of Conformity

When biased signals dominate, conformity can amplify individual errors into systematic collective failure (Han et al., 2025). We examine these failure modes in GPT-4o under **Distributed Consensus** through the confusion-matrix heatmaps (Figure 6). First, increased connectivity exposes agents to a broader set of peer signals, which attenuates individual evidence while promoting group-level alignment. Second, larger α (stronger self-reliance) generally improves accuracy, confirming that maintaining sufficient independence is essential for resisting misleading social influence. However, groups occasionally converge on false claims with high certainty, entrenching misinformation rather than correcting it. Such patterns reveal the ambivalent nature of conformity: it enhances reliability under balanced conditions, yet can solidify collective hallucinations when early biases dominate.

6 Conclusion

Our study shows that conformity in LLM-based MAS is shaped by both network topology and self-social weighting. In Centralized Aggregation, the reliability of collective outcomes is tightly coupled to hub competence, with stronger hubs amplifying overall system accuracy. In Distributed Consensus, conformity arises from iterative local interactions, with denser connectivity simultaneously enhancing convergence efficiency and average accuracy while heightening susceptibility to information cascades. While conformity facilitates consensus, it risks cementing high-confidence errors when initial biases prevail. Consequently, practical MAS design must balance efficiency and robustness through careful topology design, weighting calibration, and control of premature convergence.

Limitations

Modeling Assumptions. First, the proposed confidence-normalized pooling rule relies on self-reported confidence generated by LLMs, which is not guaranteed to be well calibrated and may partially reflect stylistic assertiveness rather than a faithful probability of correctness. Moreover, the self–social weighting parameter is fixed per run and shared across agents, and most distributed-consensus experiments restrict interaction to a fixed, structured per-round update template rather than open-ended deliberation, questioning, or tool-augmented evidence seeking.

Scale and Model Selection. Second, all experiments are conducted with a fixed group size ($N = 7$) and a limited set of three LLMs. This design is motivated by three considerations. First, many contemporary MAS deployments remain modest in size due to coordination and cost constraints. Second, fixing $N = 7$ enables strict alignment with the three-level hierarchical configuration (1 root, 2 intermediate aggregators, and 4 leaves). Third, the selected models are chosen primarily to avoid knowledge leakage arising from recent knowledge cutoffs. Our objective is not to maximize the detection accuracy, but to examine how agent-level interaction mechanisms shape the collective behavior. As a result, while stronger models may shift overall performance levels, we expect the relative interaction effects identified in this study to remain qualitatively robust.

Task and Domain Scope. Third, the study is restricted to binary misinformation detection. While this task is well suited for isolating conformity dynamics, it represents only a narrow class of collective reasoning problems. More open-ended tasks, such as multi-hop reasoning, policy analysis, or creative collaboration, may exhibit qualitatively different interaction patterns in which conformity interacts with exploration, role specialization, or strategic behavior beyond the scope of the current framework.

References

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8038–8057, Albuquerque, New Mexico. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Abdollah Amirkhani and Amir Hossein Barshooi. 2022. Consensus in multi-agent systems: a review. *Artificial Intelligence Review*, 55(5):3897–3935.

Brian DO Anderson and Mengbin Ye. 2019. Recent advances in the modelling and analysis of opinion dynamics on influence networks. *International Journal of Automation and Computing*, 16(2):129–149.

Mourad Aouini and Jinan Loubani. 2025. Towards more contextual agents: An extractor-generator optimization framework. *Preprint*, arXiv:2502.12926.

Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, Alex Lyman, Justin Olcott, Jackson Pond, and David Wingate. 2025. Testing theories of political persuasion using ai. *Proceedings of the National Academy of Sciences*, 122(18):e2412815122.

Guangyan Bao, Lifeng Ma, and Xiaojian Yi. 2022. Recent advances on cooperative control of heterogeneous multi-agent systems subject to constraints: a survey. *Systems Science & Control Engineering*, 10(1):539–551.

Sushil Bikhchandani, David Hirshleifer, Omer Tamuz, and Ivo Welch. 2024. Information cascades and social learning. *Journal of Economic Literature*, 62(3):1040–93.

Carla Capuano and Peggy Chekroun. 2024. A systematic review of research on conformity. *International Review of Social Psychology*, 37(1).

Pei Chen, Shuai Zhang, and Boran Han. 2024. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1720–1738, Mexico City, Mexico. Association for Computational Linguistics.

Li Cheng, Yijie Wang, and Xinwang Liu. 2021. Neighborhood consensus networks for unsupervised multi-view outlier detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7099–7106.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. 2025. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5123–5139, Vienna, Austria. Association for Computational Linguistics.

Pedro Cisneros-Velarde. 2025. Biases in opinion dynamics in multi-agent systems of large language models: A case study on funding allocation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1889–1916, Albuquerque, New Mexico. Association for Computational Linguistics.

Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. 2025. Understanding the uncertainty of llm explanations: A perspective based on reasoning topology. *Preprint*, arXiv:2502.17026.

Guozhu Ding, Zuer Liu, Shan Li, Jie Cao, and Zhuohai Ye. 2025. Impact of mindset types and social community compositions on opinion dynamics: A large language model-based multi-agent simulation study. *Computers in Human Behavior*, 172:108730.

Yucheng Dong, Zhaogang Ding, and Gang Kou. 2024. Social network degroot model. *Springer Books*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Arnab Kumar Ghoshal, Nabanita Das, Soham Das, and Subhankar Dhar. 2025. Minimizing spread of misinformation in social networks: a network topology based approach. *Social Network Analysis and Mining*, 15(1):15.

Florian Grötschla, Luis Müller, Jan Tönshoff, Mikhail Galkin, and Bryan Perozzi. 2025. Agentsnet: Coordination and collaborative reasoning in multi-agent llms. *Preprint*, arXiv:2507.08616.

Chen Han and Xijin Tang. 2025. An agent-based simulation framework for misinformation susceptibility test with llms: Insights from psychological factors. In *Knowledge and Systems Sciences*, pages 355–368, Singapore. Springer Nature Singapore.

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*, pages 15125–15140, Suzhou, China. Association for Computational Linguistics.

Luzie Helfmann, Nataša Djurdjevac Conrad, Philipp Lorenz-Spreen, and Christof Schütte. 2023. Modelling opinion dynamics under the impact of influencer and media strategies. *Scientific Reports*, 13:19375.

Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit. 2025. Ai–ai bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences*, 122(31):e2415697122.

Grace J. Li and Mason A. Porter. 2023. Bounded-confidence model of opinion dynamics with heterogeneous node-activity levels. *Phys. Rev. Res.*, 5:023179.

Hui Li, Ante Wang, kunquan li, Zhihao Wang, Liang Zhang, Delai Qiu, Qingsong Liu, and Jinsong Su. 2025. A multi-agent framework with automated decision rule optimization for cross-domain misinformation detection. *Preprint*, arXiv:2503.23329.

OpenAI. 2023. Gpt-3.5 turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2025-09-10.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-09-10.

Flávio L. Pinheiro and Vítor V. Vasconcelos. 2025. Heterogeneous update processes shape information cascades in social networks. *Scientific Reports*, 15(1):13999.

Mohammad Shirzadi, Emilio Cruciani, and Ahad N. Zehmakan. 2025. Opinion dynamics: A comprehensive overview. *Preprint*, arXiv:2511.00401.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025. G-safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7261–7276, Vienna, Austria. Association for Computational Linguistics.

Xinghai Wei, Tingting Yuan, Jie Yuan, Dongxiao Liu, and Xiaoming Fu. 2025. Rescom: Reward-shaped curriculum for efficient multi-agent communication learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 2153–2161, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Jiaxi Yang, Mengqi Zhang, Yiqiao Jin, Hao Chen, Qingsong Wen, Lu Lin, Yi He, Weijie Xu, James Evans, and Jindong Wang. 2025. Topological structure learning should be a research priority for llm-based multi-agent systems. *Preprint*, arXiv:2505.22467.

Shaowei Zhang and Deyi Xiong. 2025. Debate4MATH: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824, Vienna, Austria. Association for Computational Linguistics.

A Prompt Templates

A.1 Centralized Aggregation

A.1.1 Peripheral Agent (Leaf): One-shot Judgment

Use: Leaf/peripheral agent issues a single-round decision; no peer inputs are seen.

Prompt.

```
You are a careful fact-checking agent. Read the
↪  claim and produce a binary
decision and a calibrated confidence.

[Task]
1) Decide whether the claim is True (=0) or False
↪  (=1).
```

```
2) Provide a confidence p in [0, 1].
3) Give a concise justification (<100 words)
↪  grounded in the background profile
  and logical reasoning. Avoid speculation.

[Background Profile]
BACKGROUND

[Claim]
CLAIM

[Output Requirements]
Output strictly the following JSON object:

      "y": 0 or 1,      // 0=True, 1=False
      "p": number,      // float in [0,1], with
    up to 2 decimals
      "just": "..."    // <100 words, no markdown

Do NOT include any extra keys, prose, or
↪  formatting.
```

A.1.2 Hub/Root: One-shot Judgment with Submitted Leaves

Use: The hub reads the claim and sees peer JSON reports; weighting is handled by Eq. (1).

Prompt.

```
You are the central (hub/root) fact-checking
↪  agent. You will read the claim
and also see a list of peer reports. Produce your
↪  own final judgment.

[Claim]
CLAIM

[Peer Reports]
List of JSON objects from peripheral agents:
PEER_JSON_LIST
/*
Each element is:
"agent_id": "Li_i", "y": 0|1, "p": [0,1], "just":
↪  "..."
Do not copy them verbatim in your output. Use
↪  them only as additional evidence.
*/

[Task]
1) Decide whether the claim is True (=0) or False
↪  (=1).
2) Provide a calibrated confidence p in [0, 1].
3) Give a concise justification (<100 words) that
↪  references the most
  diagnostic considerations. Avoid
  ↪  majority-following; reason on merits.

[Output Requirements]
Output strictly the following JSON object:

  "y": 0 or 1,
  "p": number,
  "just": "..."

Do NOT include any extra keys, prose, or
↪  formatting.
```

A.2 Distributed Consensus

A.2.1 Initial Judgment

Use: Each agent produces its initial decision before any interaction.

Prompt.

```
You are an autonomous agent in a distributed
↪  consensus setting (no central
controller). Produce your initial judgment.

[Background Profile]
BACKGROUND

[Claim]
CLAIM

[Task]
1) Output y in 0,1 where 0=True and 1=False.
2) Output a calibrated confidence p in [0,1].
3) Provide a brief justification (<100 words)
↪  grounded in facts and logic.

[Output Requirements]
Output strictly the following JSON object:

      "agent_id": "AGENT_ID",
      "t": 0,
      "y": 0 or 1,
      "p": number,
      "just": "..."

Do NOT include any extra keys, prose, or
↪  formatting.
```

A.2.2 Interactive Rounds

Use: At each round $t \geq 1$, the agent receives its previous state and neighbor reports.

Prompt.

```
You are participating in a multi-round
↪  distributed consensus process. You will
see your previous stance and your neighbors'
↪  reports at round t-1. Provide your
current stance, but note that the final state is
↪  computed by the system.

[Context]
• Agent: AGENT_ID
• Round: ROUND
• Your previous state (t-1): "y": Y_PREV, "p":
↪  P_PREV
• Neighbor reports (t-1): NEIGHBOR_JSON_LIST

[Task]
1) State your CURRENT stance (y in 0,1, p in
↪  [0,1]).
2) Provide a compact justification (<80 words)
↪  referencing decisive signals.
3) Do not summarize all neighbors; mention only
↪  what materially changes your stance.

[Output Requirements]
```

```
Output strictly the following JSON object:

      "agent_id": "AGENT_ID",
      "t": ROUND,
      "y": 0 or 1,
      "p": number,
      "just": "..."

Do NOT include any extra keys, prose, or
↪  formatting.
```

A.3 Pseudocode for Agent Decision Process

The following pseudocode describes the agent's decision-making process and belief update mechanism:

```
# Initialize agent's state (beliefs,
    confidence)
agents = initialize_agents()

for t in range(T):  # Loop for T rounds
    of decision-making and belief
    updates
    for agent in agents:
        # Step 1: Evaluate claim using
            the prompt (Figure 2)
        claim = get_claim(t)
        background_profile =
            generate_background_profile(
            claim)
        task_description =
            generate_task_description()

        # Agent produces a binary
            judgment and confidence
            score
        y_i_t, p_i_t = evaluate_claim(
            agent, background_profile,
            task_description)

        # Step 2: Update agent's belief
            score using the update rule
            (Equation 1)
        s_i_t_plus_1 = update_belief(
            agents, agent, y_i_t, p_i_t)

        # Step 3: Assign updated belief
            back to the agent
        agent.belief_score =
            s_i_t_plus_1

# End of simulation
```

Listing 1: Pseudocode for the Agent Decision Process

B Model Settings

For the closed-source LLMs (GPT-3.5 and GPT-4o), we set the sampling temperature to 0.7 for all runs. We use `gpt-3.5-turbo-16k-0613` and `gpt-4o-1120`. For the open-source model `llama3.3-70b-instruct`, we adopt the same temperature to ensure comparability; all other decoding hyperparameters follow the default configuration of Ollama.[1] We also report the zero-shot performance of each model in Table 4.

Table 4: Zero-shot performance of LLMs on Snopes25.

Model	Acc.	Prec.	Rec.	F1
GPT-3.5	0.64	0.62	0.60	0.61
GPT-4o	0.70	0.66	0.69	0.67
Llama3.3	0.67	0.63	0.66	0.65

C Cases

We present representative case studies from the Snopes25 dataset to illustrate both successful and failure modes of our system. Cases C.1–C.4 demonstrate correct collective judgments under different topologies, while Case C.5 highlights a failure scenario in which distributed consensus amplifies an initially incorrect interpretation.

C.1 Hierarchical Centralized Aggregation (Correct)

Claim. *Pathogens can be released into the air when a toilet is flushed without a closed lid.*

Ground Truth. True.

Topology and Protocol. A three-level hierarchy with seven agents: a root (R), two intermediate aggregators (M_L, M_R), and four leaves (L1–L4). A single-round protocol is used: leaves issue one-shot judgments; intermediates summarize assigned leaves; the root produces the final decision.

Leaf Outputs ($t = 0$).

```
"agent_id":"L1","t":0,"y":0,"p":0.71,"just":"Ev
↪  idence on toilet plume aerosolization
↪  indicates airborne release without a closed
↪  lid."
"agent_id":"L2","t":0,"y":0,"p":0.68,"just":"Me
↪  chanistic fluid dynamics and observed
↪  droplet formation support potential airborne
↪  dispersion."
"agent_id":"L3","t":0,"y":0,"p":0.76,"just":"St
↪  udies show particle counts rise after
↪  flushing; lids reduce but absence increases
↪  emission."
"agent_id":"L4","t":0,"y":0,"p":0.73,"just":"Re
↪  ported bioaerosols align with the claim under
↪  open-lid flushing."
```

Intermediate Aggregators.

[1] https://github.com/ollama/ollama

```
"y":0,"p":0.82,"just":"Leaf reports consistently
↪  indicate aerosolized particles after
↪  flushing without a lid."
"y":0,"p":0.80,"just":"Multiple leaves cite
↪  increased particle counts and bioaerosols;
↪  the claim is supported."
```


Root Decision.


```
"y":0,"p":0.86,"just":"Both sub-aggregators
↪  converge on airborne release under open-lid
↪  flushing."
```


C.2 Star Centralized Aggregation (Correct)

Claim. *Pathogens can be released into the air when a toilet is flushed without a closed lid.*

Ground Truth. True.

Topology and Protocol. A star topology with one hub (H) and six leaves (L1–L6). Leaves issue one-shot judgments; the hub produces the final decision.

Leaf Outputs ($t = 0$).


```
"agent_id":"L1","t":0,"y":0,"p":0.72,"just":"Op
↪  en-lid flushing generates toilet plumes with
↪  aerosolized particles."
"agent_id":"L2","t":0,"y":0,"p":0.69,"just":"Dr
↪  oplet and aerosol formation supports
↪  airborne release without a lid."
"agent_id":"L3","t":0,"y":0,"p":0.75,"just":"Pa
↪  rticle counts increase after flushing; lids
↪  mitigate emissions."
"agent_id":"L4","t":0,"y":0,"p":0.71,"just":"Bi
↪  oaerosol evidence aligns with open-lid
↪  flushing."
"agent_id":"L5","t":0,"y":0,"p":0.74,"just":"Fl
↪  uid dynamics indicate an upward plume capable
↪  of suspending microbes."
"agent_id":"L6","t":0,"y":0,"p":0.70,"just":"Ob
↪  served plume height supports airborne
↪  dispersal."
```


Hub Decision.


```
"y":0,"p":0.86,"just":"All leaves converge on
↪  plume and aerosol evidence; the claim is
↪  true."
```


C.3 Ring Distributed Consensus (2 Neighbors, Correct)

Claim. *Pathogens can be released into the air when a toilet is flushed without a closed lid.*

Ground Truth. True.

Topology and Outcome. A ring of seven agents, each connected to two neighbors. The system reaches unanimous consensus ($y = 0$) at round $t = 3$.

Initial Judgments ($t = 0$).


```
"agent_id":"A1","t":0,"y":0,"p":0.62,"just":"To
↪  ilet plume studies indicate aerosol
↪  release."
"agent_id":"A2","t":0,"y":1,"p":0.58,"just":"Ev
↪  idence appears
↪  mixed."
"agent_id":"A3","t":0,"y":0,"p":0.65,"just":"Pa
↪  rticle counts increase after
↪  flushing."
"agent_id":"A4","t":0,"y":1,"p":0.55,"just":"Ef
↪  fect size may be
↪  limited."
"agent_id":"A5","t":0,"y":0,"p":0.60,"just":"Fl
↪  uid dynamics support upward plume
↪  formation."
"agent_id":"A6","t":0,"y":0,"p":0.63,"just":"Bi
↪  oaerosol reports align with open-lid
↪  flushing."
"agent_id":"A7","t":0,"y":1,"p":0.57,"just":"Pr
↪  ior evidence seems
↪  inconclusive."
```


Across rounds $t = 1$ and $t = 2$, local neighbor interactions gradually shift initially skeptical agents toward the majority stance.

Consensus ($t = 3$).


```
"agent_id":"A1","t":3,"y":0,"p":0.74,"just":"Ne
↪  ighborhood fully aligned on plume
↪  evidence."
"agent_id":"A2","t":3,"y":0,"p":0.73,"just":"Su
↪  stained agreement justifies
↪  True."
"agent_id":"A3","t":3,"y":0,"p":0.75,"just":"Ev
↪  idence remains consistent across
↪  rounds."
"agent_id":"A4","t":3,"y":0,"p":0.70,"just":"Ne
↪  ighbor data resolves prior
↪  uncertainty."
"agent_id":"A5","t":3,"y":0,"p":0.74,"just":"Co
↪  nvergent aerosol
↪  observations."
"agent_id":"A6","t":3,"y":0,"p":0.75,"just":"Su
↪  pportive studies maintain
↪  True."
"agent_id":"A7","t":3,"y":0,"p":0.73,"just":"Cu
↪  mulative local evidence confirms the
↪  claim."
```


C.4 Complete Graph Distributed Consensus (6 Neighbors, Failure Case)

Claim. *When spiders sense danger, they run toward people for protection.*

Ground Truth. False.

Topology and Outcome. A complete graph with seven agents (six neighbors each). Despite initial disagreement, the system reaches unanimous consensus ($y = 0$, True) at round $t = 3$, resulting in a confident but incorrect judgment.

Initial Judgments ($t = 0$).


```
"agent_id":"A1","t":0,"y":1,"p":0.63,"just":"Sp
↪ iders generally avoid humans; approach seems
↪ unlikely."
"agent_id":"A2","t":0,"y":0,"p":0.58,"just":"So
↪ me anecdotal accounts suggest refuge near
↪ large objects."
"agent_id":"A3","t":0,"y":1,"p":0.61,"just":"Ty
↪ pical response is retreat from vibration
↪ sources."
"agent_id":"A4","t":0,"y":0,"p":0.55,"just":"Mo
↪ vement toward stationary masses could reduce
↪ exposure."
"agent_id":"A5","t":0,"y":1,"p":0.60,"just":"Et
↪ hology literature emphasizes avoidance
↪ behavior."
"agent_id":"A6","t":0,"y":0,"p":0.57,"just":"Sh
↪ elter-seeking may incidentally align with
↪ human location."
"agent_id":"A7","t":0,"y":1,"p":0.59,"just":"Av
↪ ailable cues suggest retreat, not
↪ approach."
```


Through rounds $t = 1$ and $t = 2$, repeated exposure to a plausible but weak "nearest-cover" interpretation increases its perceived credibility, despite limited empirical support.

Erroneous Consensus ($t = 3$).


```
"agent_id":"A1","t":3,"y":0,"p":0.80,"just":"Ne
↪ arest-cover account explains apparent
↪ approach behavior."
"agent_id":"A2","t":3,"y":0,"p":0.72,"just":"Co
↪ nvergent explanations support
↪ True."
"agent_id":"A3","t":3,"y":0,"p":0.71,"just":"Co
↪ nsistent shelter-seeking behavior validates
↪ interpretation."
"agent_id":"A4","t":3,"y":0,"p":0.74,"just":"Ai
↪ rflow and vibration gradients provide
↪ plausible refuge."
"agent_id":"A5","t":3,"y":0,"p":0.79,"just":"Ne
↪ arest-cover heuristic yields apparent
↪ protection-seeking."
"agent_id":"A6","t":3,"y":0,"p":0.75,"just":"Ne
↪ twork agreement justifies the
↪ conclusion."
"agent_id":"A7","t":3,"y":0,"p":0.73,"just":"Cu
↪ mulative reports support
↪ approach-for-protection behavior."
```


This case illustrates how strong conformity pressure in highly connected networks can amplify a coherent yet incorrect narrative, leading to confident misclassification.

D Formal Definition of Metrics

D.1 Centralized Aggregation Metrics

Let $y^\star \in \{0, 1\}$ denote the ground-truth label for a given claim, and let y_c denote the final judgment issued by the central node (hub or root).

Central Accuracy (CA). Central Accuracy measures the correctness of the collective decision produced by the central node:

$$\mathrm{CA} = \mathbb{I}[y_c = y^\star], \tag{3}$$

where $\mathbb{I}[\cdot]$ is the indicator function. CA directly reflects the reliability of hub-mediated aggregation and isolates the effect of central competence from peripheral diversity.

Peripheral Accuracy (PA). Let $\mathcal{P}$ denote the set of peripheral (non-central) agents. Peripheral Accuracy is defined as

$$\mathrm{PA} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbb{I}[y_i = y^\star], \tag{4}$$

capturing the average individual-level correctness independent of the aggregation outcome.

Center–Periphery Consistency (CPC). To quantify immediate conformity to the central decision, we define

$$\mathrm{CPC} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbb{I}[y_i = y_c]. \tag{5}$$

CPC measures the extent to which peripheral agents align with the hub's judgment, irrespective of correctness, thereby isolating conformity strength from accuracy.

D.2 Distributed Consensus Metrics

In Distributed Consensus, agents iteratively update their judgments over rounds $t = 0, \ldots, T_{\max}$. Let $y_i^{(t)}$ denote agent i's judgment at round t.

Final Accuracy (FA). Let $\hat{y}$ denote the group decision at termination, defined as the unanimous label if consensus is reached, or the majority label at $T_{\max}$ otherwise. Final Accuracy is given by

$$\mathrm{FA} = \mathbb{I}[\hat{y} = y^\star], \tag{6}$$

evaluating whether the emergent collective outcome matches the ground truth.

Time-to-Consensus (TTC). Time-to-Consensus is defined as

$$\mathrm{TTC} = \min\left\{t : y_1^{(t)} = y_2^{(t)} = \cdots = y_N^{(t)}\right\}, \tag{7}$$

and is undefined if unanimity is not achieved within $T_{\max}$. TTC reflects the efficiency of convergence induced by network connectivity and conformity strength.

Conformity Index (CI). At each round t, the Conformity Index is defined as

$$\mathrm{CI}^{(t)} = \max\left(\begin{array}{l}\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathbb{I}\left[y_i^{(t)} = 1\right], \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathbb{I}\left[y_i^{(t)} = 0\right]\end{array}\right) \tag{8}$$

For binary judgments, $\mathrm{CI}^{(t)} \in [0.5, 1]$, where $\mathrm{CI}^{(t)} = 1$ indicates unanimity and $\mathrm{CI}^{(t)} = 0.5$ corresponds to maximal disagreement.

Average Conformity Index (ACI). To summarize conformity dynamics over time, we report the Average CI:

$$\mathrm{ACI} = \frac{1}{T_{\max}}\sum_{t=1}^{T_{\max}}\mathrm{CI}^{(t)}, \tag{9}$$

which captures the overall tendency toward alignment across interaction rounds, rather than only the terminal state.