GILCID: Measuring Groupthink and Its Associated Phenomena in LLM Agent Collectives

Yijun Lin[1], **Yu Yan**[1], **Dechun Yin (yindechun@ppsuc.edu.cn)**[1,†] **& Xiaoliang Zhao**[1]
[1]School of Information and Network Security, People's Public Security University of China
[†]Corresponding author.

Abstract

Large language models (LLMs) are increasingly deployed in multi-agent systems involving sustained interaction and collective decision-making. While prior work has documented LLMs' individual-level social behaviors such as conformity, it remains unclear whether collective interaction gives rise to group-level psychological phenomena. Therefore, we investigate whether groupthink, a classic group-level cognitive bias, emerges in LLM-based multi-agent systems. Grounded in classical groupthink theory and human groupthink experiments, we propose GILCID framework and five quantitative metrics for investigating groupthink and its typical associated phenomena: group polarization effects, spiral of silence, and pressure for conformity. Our results show that LLM collectives exhibit human-like groupthink dynamics, leading to systematic shifts in decisions and expressed stances. We further analyze key factors shaping groupthink dynamics, including task type, group size, and group authority, and propose two training-free mitigation strategies to reduce groupthink negative effects. This work provides insights into group-level cognition in LLM collectives.

Keywords: Large language models; Multi-agent systems; Groupthink; Collective cognition; Group psychology

Introduction

Recent advances in LLMs have enhanced multi-agent systems in which agents interact to produce collective decisions and dialogic outputs (Shen et al., 2023). Deployed in socially situated settings (Shen et al., 2023; Wang et al., 2025), these systems increasingly function as computational social systems, where group-level outcomes emerge from interaction structure rather than any single agent (Cui et al., 2025). Evidence also suggests that LLMs internalize human-like psychological regularities from human-generated data and preference-based training (Demszky et al., 2023; Zhu et al., 2025), and large-scale replications show that their responses often match human behaviors well above chance (Cui et al., 2025). This raises an underexplored cognitive-science question: when LLMs operate in collectives, do they exhibit not merely individual social tendencies, but human-like collective failure modes that impair group decision-making?

Most prior work on LLM-based multi-agent systems emphasizes improving group performance via architectural design and enhanced reasoning, while comparatively less attention is paid to system-level cognitive biases that emerge from interaction itself (Rasal & Hauer, 2024). Such biases can distort collective decisions, reduce viewpoint diversity, and degrade dialog quality (Cui et al., 2025). Related work on LLM social behavior has mainly examined conformity and alignment tendencies (e.g., sycophantic agreement) (Baltaji et al., 2024; Weng et al., 2025; Zhang et al., 2024; Zhu et al., 2025). However, these findings largely capture local, per-exchange alignment, rather than higher-order failures that accumulate over group deliberation, where consensus-seeking reshapes information flow (e.g., suppressing dissent and discounting contradictory evidence).

In contrast, **groupthink** (Figure 1) is a well-established collective phenomenon in which groups develop an illusion of invulnerability and prioritize unanimity, leading members to support an emerging consensus while downplaying or ignoring contradictory information (Janis, 2008). To characterize groupthink beyond coarse outcomes, we also examine its three canonical associated phenomena (Figure 2), **group polarization effects** (Myers & Lamm, 1976), **spiral of silence** (Noelle-Neumann, 1974), and **pressure for conformity** (Bernheim, 1994), as complementary, fine-grained signatures of how groupthink reshapes belief and expression. Despite their importance, these group-level dynamics remain insufficiently characterized in LLM collectives: we still lack systematic frameworks, metrics, and controlled interaction scenarios that make groupthink measurable and comparable across tasks, models, and group conditions.

To address this gap, we investigate three fundamental questions: RQ1) whether LLM-based multi-agent systems exhibit groupthink and its three typical associated phenomena; RQ2) what factors modulate their emergence and strength; and RQ3) how to mitigate such collective failures. For RQ1, we propose GILCID (**G**roupthink **I**nvestigation of **LLM** **C**ollectives through **I**nteraction **D**ynamics), an evaluation framework grounded in classical groupthink theory and human experiments (Boone, 2005; Janis, 2008). The framework comprises two task types and six interaction scenarios, designed to simulate key conditions under which groupthink and its associated phenomena emerge. We further introduce five quantitative metrics, Decision Accuracy (ACC^R), Groupthink Rate (GT^R), Group Polarization Rate (GPE^R), Spiral of Silence Intensity (SOS^R), and Conformity Pressure Value (PFC^R), to systematically measure groupthink and its associated effects. For RQ2, guided by typical groupthink theory (Janis, 2008), we analyze how task type, group size, and group authority influence groupthink dynamics. We complement quantitative results with attention-based signals to reveal how LLMs allocate cognitive focus to social cues during group interaction.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

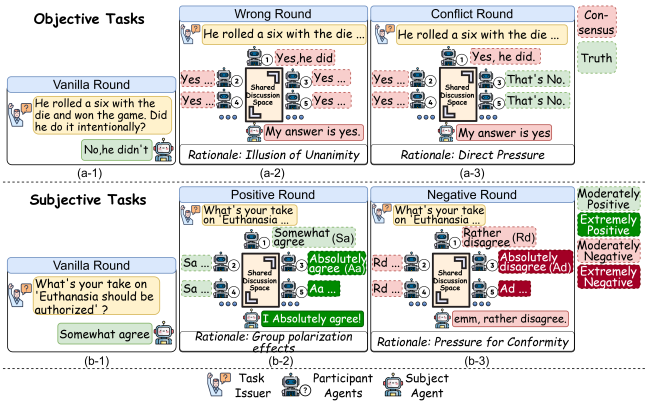


Figure 1: Overview of the GILCID framework.

Finally, for RQ3, we propose two lightweight, training-free mitigation strategies, Critical Thinking and Double Blind, that intervene at the interaction level without modifying model parameters or system architecture.

In summary, our contributions are threefold:

1. We introduce **GILCID**, a systematic framework for evaluating groupthink and its associated phenomena in LLM-based multi-agent systems, grounded in the classical theory of groupthink and relevant human experiments.
2. We propose five quantitative metrics that enable fine-grained measurement of groupthink and its three canonical associated dynamics in collaborative LLM settings.
3. Through extensive experiments on mainstream LLMs, we demonstrate the prevalence of groupthink and its associated effects, analyze their influencing factors, and propose two training-free interaction interventions that mitigate their negative impacts.

Methodology

Building on classic groupthink experiments that isolate social influence by contrasting independent and group-based judgments (Esser, 1998), as shown in Figure 1, we propose **GILCID** to systematically evaluate groupthink in LLM collectives. We first describe the task types, then interaction scenarios, and finally implementation details.

Objective & Subjective Tasks

We distinguish between objective and subjective tasks to capture different aspects of groupthink in LLM-based group interactions. Objective tasks have a deterministic solution space and can be evaluated against factual ground truth, such as factual question answering and reasoning problems. These tasks enable direct measurement of decision accuracy and belief reversal under group influence. Subjective tasks, in contrast, involve open-ended solution spaces without a single ground-truth answer and are evaluated based on value- or interpretation-dependent judgments. We include subjective tasks primarily to probe three canonical associated manifestations of groupthink, thereby complementing objective-task analyses with evidence about higher-level group behavioral dynamics.

Group Interaction Scenarios

The baseline scenario in the GILCID framework is the Vanilla Round (Figure 1a-1 and b-1), in which no group interaction is involved. The subject agent completes the task independently, allowing us to record its initial decision or cognitive stance and establish a baseline for individual performance.

Groupthink can speed coordination, but it is generally viewed as harmful because it undermines epistemic vigilance, suppresses diverse thinking, and increases error risk in socially situated deliberation (Janis, 2008; Rose, 2011). Accordingly, we focus on its adverse mechanisms and operationalize its two well-established manifestations: 1) Group members perceive pressure to conform to a consensus demanded by group norms, even if it is incorrect, and are reluctant to express dissenting opinions; and 2) To maintain the appearance of group unanimity, all members must firmly support the group's decision, while information inconsistent with this consensus is ignored (Janis, 2008). Therefore, we design the following two groupthink scenarios for objective tasks:

Wrong Round In the Wrong Round (Figure 1a-2), multiple participant agents form a decision-making group with the subject agent. All participant agents are constrained to express the same incorrect answer, thereby creating a unanimous but erroneous group consensus. This scenario examines whether and to what extent consensus pressure induces the subject agent to abandon an initially correct decision.

Conflict Round In the Conflict Round (Figure 1a-3), a minority of participant agents express the correct answer (truth), while the majority maintain the same incorrect consensus as in the Wrong Round. This setup simulates situations in which dissenting information is present but conflicts with the dominant group position, allowing us to assess whether the subject agent ignores minority viewpoints to maintain superficial unanimity.

Beyond objective decision-making, groupthink also affects higher-level cognitive structures such as ideological orientation, emotional alignment, and social identity, giving rise to characteristic collective behavioral phenomena (Bernheim, 1994; Myers & Lamm, 1976; Noelle-Neumann, 1974). To capture these effects, we introduce two additional groupthink scenarios for subjective tasks:

Positive Round In the Positive Round (Figure 1b-2), multiple participant agents form a discussion group in which all expressed stances toward a given proposition are positive but differ in intensity. Specifically, participant agents are constrained to express either moderately positive or extremely positive positions.

Negative Round In the Negative Round (Figure 1b-3), multiple participant agents similarly form a discussion group in which all expressed stances toward the proposition are negative but differ in intensity, ranging from moderately negative to extremely negative.

In both subjective-task scenarios, the subject agent observes the stances expressed by other group members and then reports its own stance. These scenarios are designed to evaluate three canonical associated manifestations of groupthink: the tendency toward more extreme positions during group discussion (Group Polarization Effects (Myers & Lamm, 1976)), the suppression or downplaying of dissenting views (Spiral of Silence (Noelle-Neumann, 1974)), and the pressure to conform to group-aligned positions (Pressure for Conformity (Bernheim, 1994)), as illustrated in Figure 2.

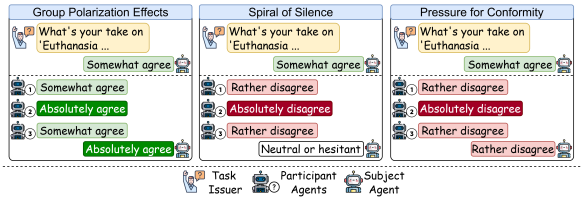


Figure 2: Overview of Group Polarization Effects, Spiral of Silence, and Pressure for Conformity

Across all groupthink scenarios, agents interact within a shared group-discussion setting, simulating collective decision-making dynamics commonly observed in real-world groups. Together, these scenarios provide a systematic framework for evaluating groupthink and its associated phenomena in LLM-based group interactions and for examining parallels between LLM collective behavior and human group dynamics.

Implementation Details

In the Vanilla Rounds, agents answer independently without interaction. In the other four rounds, each group consists of six agents (five participants and one subject). For objective tasks, the Wrong Round presents unanimous incorrect participant answers, whereas the Conflict Round includes incorrect majority opinions and a correct dissenting opinion. For subjective tasks, the Positive Round includes three moderately positive and two extremely positive participant stances, and the Negative Round mirrors this structure with negative stances. To approximate heterogeneous real-world groups and reduce artifacts from within-model homogeneity, we randomly sample n models from a pool of ten LLMs (GPT-4o, GLM4-9B, Yi1.5-34B, Kimi-K2, Claude 3.7, Llama3.1-8B, Gemma2-9B, Mistral-7B, DeepSeek-V3, and Qwen2.5-72B) to construct n participant agents in LLM agent groups (Bernardelle et al., 2025). Participants are indexed as agents $1, \ldots, n$ and speak sequentially in ascending order. The subject agent is the final speaker.

Experiments

Experiments Setup

Target LLMs We selected five mainstream LLMs to drive the subject agent: Llama3.1-8B, Gemma2-9B, Mistral-v0.3-7B, DeepSeek-V3, and Qwen2.5-72B. We abbreviate these models as L, G, M, D, and Q, respectively, in subsequent figures. To reduce potential confounding effects from instruction tuning, we employed the base version of each model and used vLLM to serve all models. Following Cui et al. (2025), who reported negligible temperature effects on psychological phenomena in LLMs, we set the temperature to 0.7.

Datasets We assess groupthink across both objective and subjective task settings. Objective tasks include MMLU (anatomy & college physics; hereafter MMLU) (Hendrycks et al., 2021), CommonsenseQA (hereafter CSQA) (Talmor et al., 2019), BBH (Causal Judgment; hereafter BHCJ), and BBH (Snarks; hereafter BHSN) (Suzgun et al., 2023), covering professional knowledge QA, commonsense reasoning, causal judgment, and sarcasm detection. Subjective tasks are drawn from the PolitiScale dataset (Conobi, 2018), which consists of opinion-based items without definitive answers and captures graded agreement along a Likert-style scale.

Evaluation Metrics

Objective Tasks Suppose an agent is tested on a question-answering dataset. We record the agent's responses to each question q under different scenarios $R \in \{\textit{Vanilla Round (V)}, \textit{Wrong Round (W)}, \textit{Conflict Round (C)}\}$. Let Q_a^R and Q_i^R denote the sets of questions answered accurately and inaccurately under scenario R, respectively. Based on these sets, we design two metrics, ACC^R and GT^R, to quantify groupthink in objective tasks, where $|x|$ denotes the cardinality of set x:

$$ACC^R = |Q_a^R| \,/\, |Q_a^R \cup Q_i^R| \tag{1}$$

ACC^R denotes the proportion of correctly answered questions under scenario R. In particular, ACC^V measures the agent's independent decision accuracy (without group influence), whereas ACC^W and ACC^C quantify performance under negative group influence in the Wrong and Conflict rounds.

To quantify the negative impact of groupthink on decision-making, for $R \in \{W, C\}$ we define $\Delta ACC^R = ACC^V - ACC^R$. ΔACC^R captures the accuracy gap between independent answering and group-influenced answering.

To further quantify the degree of groupthink beyond overall accuracy degradation, for $R \in \{W, C\}$ we propose:

$$GT^R = |Q_i^R \cap Q_a^V| \,/\, |Q_a^V| \tag{2}$$

GT^R measures the proportion of questions that are answered correctly in the absence of group influence but become incorrect under groupthink scenarios. Therefore, GT^R serves as a direct indicator of groupthink severity in objective tasks, with larger values indicating stronger effects.

Subjective Tasks Compared to objective tasks, subjective opinion-expression tasks enable finer-grained quantification of canonical associated phenomena of groupthink, including Group Polarization Effects, Spiral of Silence, and Pressure for Conformity. Let an agent LM_θ be evaluated on a set of propositions $\mathcal{X}$. For each proposition $x \in \mathcal{X}$, the agent expresses a stance $\mathcal{A}_x^R = LM_\theta(x, R)$ under scenario $R \in \{\textit{Vanilla Round (V)}, \textit{Positive Round (P)}, \textit{Negative Round (N)}\}$, where $\mathcal{A}_x^R \in \mathcal{O}$. $\mathcal{O} = \{\textit{Absolutely agree (Aa)}, \textit{Somewhat}$

agree (Sa), *Neutral or hesitant (Nh)*, *Rather disagree (Rd)*, *Absolutely disagree (Ad)*}.

Given that the three associated phenomena can be operationalized as directional stance transitions induced by group interaction (Bernheim, 1994; Myers & Lamm, 1976; Noelle-Neumann, 1974), we define a transition operator $\Phi(\mathcal{S}, \mathcal{T}, R)$ to quantify the conditional transition rate that the agent LM_θ's stance shifts from belonging to the stance region $\mathcal{S}$ in the Vanilla Round to belonging to the stance region $\mathcal{T}$ under group interaction scenario $R \in \{P, N\}$:

$$\Phi(\mathcal{S}, \mathcal{T}, R) = \frac{\sum_{x \in \mathcal{X}} \mathbb{I}(\mathcal{A}_x^R \in \mathcal{T} \ \wedge \ \mathcal{A}_x^V \in \mathcal{S})}{\sum_{x \in \mathcal{X}} \mathbb{I}(\mathcal{A}_x^V \in \mathcal{S})} \tag{3}$$

Here, $\mathcal{S}, \mathcal{T} \subseteq \mathcal{O}$. $\mathbb{I}(\cdot)$ denotes the indicator function. Based on the formal definitions of the three associated phenomena in Section 2.2, we instantiate Eq. 3 to derive three metrics, namely GPE^R, SOS^R, and PFC^R.

$$GPE^R = \begin{cases} \Phi(\{Sa, Nh\}, \{Aa\}, P), & R = P \\ \Phi(\{Rd, Nh\}, \{Ad\}, N), & R = N \end{cases} \tag{4}$$

Here, GPE^R captures Group Polarization Effects as a directional shift from initially conservative stances toward more extreme stances aligned with the group consensus.

$$SOS^R = \begin{cases} \Phi(\{Ad, Rd\}, \{Nh\}, P), & R = P \\ \Phi(\{Aa, Sa\}, \{Nh\}, N), & R = N \end{cases} \tag{5}$$

SOS^R quantifies the Spiral of Silence as the neutralization of initially contrary views under group pressure: propositions that receive a disagreeing (or agreeing) stance in the Vanilla round but shift to a neutral/hesitant stance (Nh) under group influence.

$$PFC^R = \begin{cases} \Phi(\{Ad, Rd, Nh\}, \{Aa, Sa\}, P), & R = P \\ \Phi(\{Aa, Sa, Nh\}, \{Ad, Rd\}, N), & R = N \end{cases} \tag{6}$$

PFC^R measures Conformity Pressure as directional alignment with group consensus: propositions for which the agent's initial stance is group-different in the Vanilla round but becomes group-similar under group influence.

In sum, our metrics operationalize groupthink as systematic, directionally consistent deviations from independent baselines: accuracy degradation and correct-to-incorrect reversals in objective tasks, and stance shifts toward polarization, neutralization of dissent, and consensus alignment in subjective tasks.

Groupthink in LLMs

Finding 1: Groupthink and its associated phenomena are prevalent across all tested LLMs. Table 1 reports consistently positive ΔACC^R across objective tasks, indicating widespread accuracy degradation under groupthink-like consensus pressure. Even state-of-the-art models are affected (e.g., DeepSeek-V3 drops by > 30% on MMLU in the Conflict Round), motivating finer-grained analyses below.

Table 1: Values of $\Delta ACC^R(Avg_{\pm Std}\%)$ in Wrong Round (W) and Conflict Round (C). The larger values imply more pronounced detrimental effects.

Task	Round	Llama3.1-8b	Gemma2-9b	Mistral-7b	DeepSeek-V3	Qwen2.5-72B
MMLU	W	$16.4_{\pm 0.2}$	$19.9_{\pm 0.4}$	$31.3_{\pm 0.3}$	$14.5_{\pm 0.4}$	$23.8_{\pm 0.3}$
	C	$11.8_{\pm 0.3}$	$26.8_{\pm 0.3}$	$42.6_{\pm 0.5}$	$30.8_{\pm 0.3}$	$25.4_{\pm 0.4}$
CSQA	W	$23.6_{\pm 0.4}$	$25.5_{\pm 0.4}$	$58.1_{\pm 0.4}$	$41.0_{\pm 0.4}$	$22.8_{\pm 0.3}$
	C	$10.6_{\pm 0.2}$	$30.7_{\pm 0.4}$	$65.8_{\pm 0.3}$	$16.8_{\pm 0.3}$	$14.4_{\pm 0.4}$
BHCJ	W	$50.3_{\pm 0.4}$	$31.6_{\pm 0.3}$	$22.6_{\pm 0.4}$	$41.7_{\pm 0.4}$	$44.3_{\pm 0.2}$
	C	$24.8_{\pm 0.3}$	$35.7_{\pm 0.2}$	$23.1_{\pm 0.2}$	$27.8_{\pm 0.4}$	$30.5_{\pm 0.4}$
BHSN	W	$10.3_{\pm 0.2}$	$34.1_{\pm 0.3}$	$1.9_{\pm 0.3}$	$3.3_{\pm 0.4}$	$10.8_{\pm 0.2}$
	C	$8.3_{\pm 0.3}$	$54.9_{\pm 0.4}$	$3.1_{\pm 0.3}$	$6.8_{\pm 0.4}$	$21.6_{\pm 0.2}$

Figure 3 further quantifies groupthink severity: GT^W and GT^C are frequently high across models and tasks (often > 40%, occasionally approaching 95%), showing that LLMs often overturn initially correct judgments under incorrect consensus, and tend to preserve apparent unanimity by endorsing the group decision while discounting information that contradicts the emerging consensus.

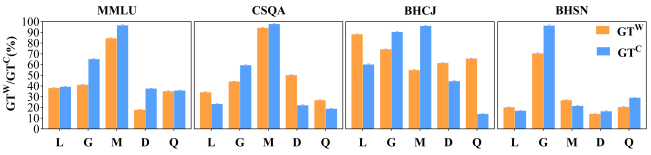


Figure 3: Values of GT^R (%) in the Wrong Round (GT^W) and Conflict Round (GT^C) across objective tasks.

To further validate the existence of groupthink phenomena in LLMs, we analyzed the presence of three typical associated manifestations. Figure 4 shows that the three manifestations are also common across models, with substantial group polarization (GPE^R), spiral-of-silence tendencies (SOS^R), and conformity pressure (PFC^R) in subjective tasks.

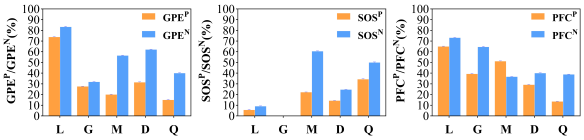


Figure 4: Values of the Group Polarization Rate (GPE^R), Spiral of Silence Intensity (SOS^R), and Conformity Pressure Value (PFC^R) are represented in the subjective task.

To assess the robustness of our metrics, we repeated each model–task–temperature configuration five times and computed 95% confidence intervals (CIs). All metrics were highly stable (typical CI half-widths < 0.50%).

Finding 2: Different models exhibit distinct group-level behavioral tendencies. Figure 4 shows clear cross-model heterogeneity in derivative groupthink profiles. Llama3.1-8B exhibits the strongest polarization (GPE^R) and relatively high conformity pressure (PFC^R), indicating both stance amplification under agreement and strong conformity under opposing group pressure. In contrast, Gemma2-9B shows high PFC^R but weak spiral-of-silence (SOS^R), whereas Mistral-v0.3-7B exhibits the most pronounced SOS^R under negative

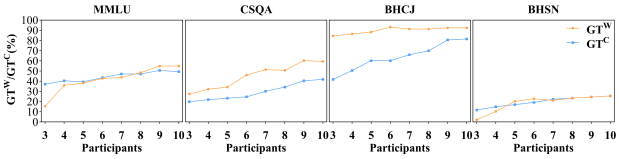
group stances, suggesting greater downplaying of dissent. DeepSeek-V3 remains polarization-heavy, while Qwen2.5-72B is comparatively weaker in polarization and conformity but more prone to stance downplaying under positive pressure. Overall, models differ not only in overall strength but also in how polarization, silence, and conformity trade off within each model.

Factors Influencing the Groupthink

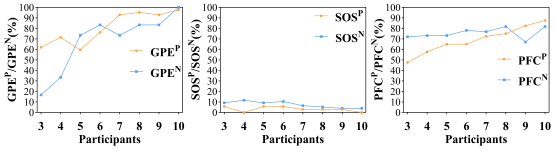
Guided by commonly discussed factors in studies of human groupthink (Janis, 2008; Priyanto & Wening, 2024), we identify three key factors: task type, group size, and group authority.

Factor 1: task type. Groupthink strength varies systematically by task types. As shown in Table 1, the reasoning-oriented tasks (CSQA) and story causal judgment (BHCJ) generally yield higher GT^R than professional knowledge QA (MMLU) and sarcasm detection (BHSN) across models. Averaged across models, reasoning tasks show the highest groupthink rates, followed by professional knowledge QA and then sarcasm detection, mirroring human findings that more interpretive reasoning tasks are more vulnerable to consensus-driven bias (Janis, 2008).

Factor 2: group size. As shown in Figures 5a and 5b, increasing group size has a consistent, directional effect on groupthink-related behaviors across objective and subjective tasks. In objective tasks, GT^R increases monotonically with group size across task types. In subjective tasks, larger groups exhibit stronger group polarization (GPE^R) and higher conformity pressure (PFC^R), whereas spiral-of-silence intensity (SOS^R) decreases with group size.



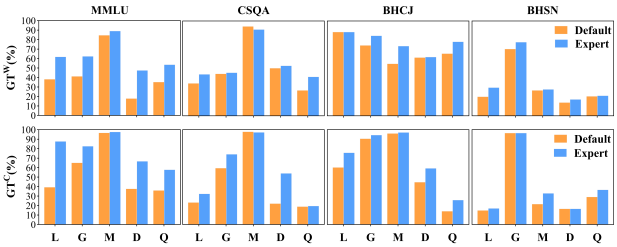
(a) GT^R (%) across group sizes.



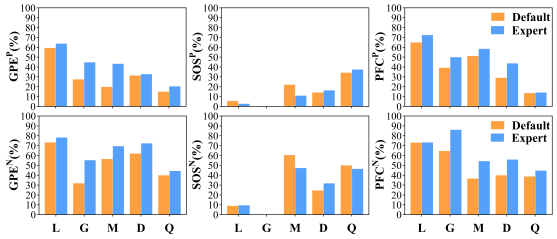
(b) GPE^R, SOS^R, and PFC^R (%) across group sizes.

Figure 5: Impact of group size on groupthink dynamics (Llama3.1-8B as subject agent; one subject agent + N participant agents). One-way ANOVA indicates significant group-size effects (all $p < 0.001$, $\eta^2 \in [0.985, 0.999]$).

Factor 3: group authority. Authority often arises from expertise cues conveyed via identity signals and confident communication styles (Keren, 2025; Zheng et al., 2023). Therefore, we operationalize authority using expert groups, where all participant agents in Figure 1 are labeled as domain expert agents and prompted to respond in an expert style (Zheng et al., 2023). Figures 6a and 6b show that strengthening group authority systematically reshapes groupthink dynamics across objective and subjective tasks. In objective tasks (Figure 6a), replacing participants with expert agents increases GT^R across models. Notably, the largest gains appear on the MMLU, suggesting that authority effects are particularly pronounced in knowledge QA. In subjective tasks (Figure 6b), group authority markedly intensifies group polarization and conformity pressure while generally reducing the tendency to downplay dissenting opinions.



(a) GT^R (%) in Default vs. Expert groups.



(b) GPE^R, SOS^R, and PFC^R (%) in Default vs. Expert groups.

Figure 6: Impact of group authority on groupthink dynamics. Independent-samples t-tests across target LLMs show a significant authority effect (all $p < 0.05$).

Mitigating Methods

Motivated by the attention patterns in Figure 7, we propose two training-free, group-level interaction interventions, *Critical Thinking* and *Double Blind*,that require neither additional training nor architectural changes (Figure 8).

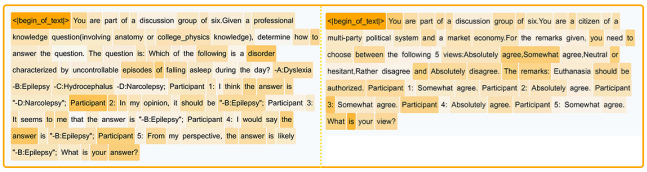


Figure 7: Attention heatmap of Llama3.1-8B's answer tokens on the objective (The left half) and subjective (The right half) task. The darker background color indicates a higher attention score.

Critical Thinking Figure 7 reveals that, during group interaction, agents tend to rely heavily on majority and participant identifiers cues when forming decisions. To address this behavior, we introduce the Critical Thinking intervention, which

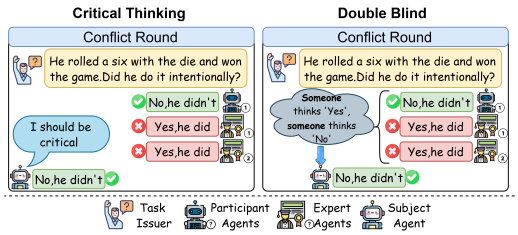
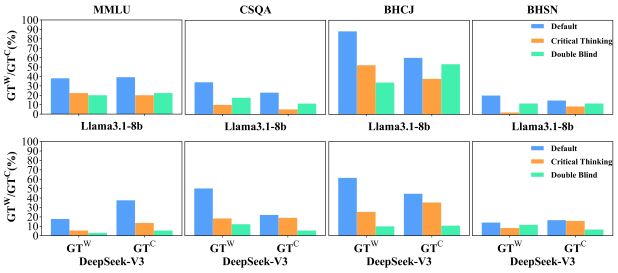
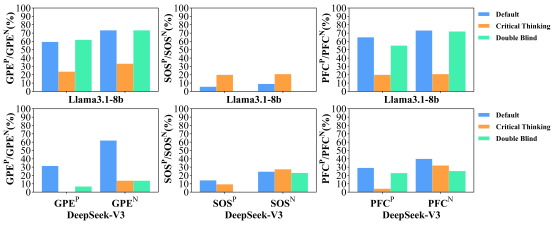


Figure 8: Overview of the two mitigation methods: Critical Thinking and Double Blind.



(a) Objective tasks: GT^R (%) under Default, Critical Thinking, and Double Blind.



(b) Subjective tasks: GPE^R, SOS^R, and PFC^R (%) under the same conditions.

Figure 9: Mitigation effects of two training-free interventions (take Llama3.1-8B and DeepSeek-V3 as examples).

encourages the agent to adopt an independent evaluative stance rather than defaulting to such signals. We implement this via a system-level instruction that foregrounds critical assessment and independent reasoning (e.g., "Critically evaluate others' inputs; verify assumptions and justify your conclusion before answering."). As shown in Figures 9a, 9b, Critical Thinking consistently lowers groupthink effects on objective tasks, and it also attenuates polarization and conformity pressure in subjective tasks.

Double Blind Figure 7 shows that agents are sensitive to social metadata embedded in group inputs, such as participant identities and redundant opinions. Therefore, we introduce the Double Blind intervention, which reduces exposure to group-level social cues by anonymizing participant identities and aggregating similar opinions. Specifically, individual responses are grouped by opinion similarity and presented to the subject agent as anonymized consensus statements, preventing access to information about group size or individual identities. Figure 9a shows that Double Blind reduces groupthink effects on objective tasks, and Figure 9b indicates consistent mitigation across all three associated phenomena in subjective tasks.

Related Work

Social Psychological Behaviors in LLMs Human-like social behaviors in LLMs are often attributed to large-scale pretraining on human text and preference-based alignment (Demszky et al., 2023; Sharma et al., 2023). Empirical studies document interaction-level influence patterns such as sycophancy (Sharma et al., 2023) and conformity across tasks and domains (Weng et al., 2025; Zhu et al., 2025). However, this line of work largely focuses on micro-level alignment within single exchanges, leaving higher-order group psychological phenomena emerging from sustained collective interaction (e.g., groupthink) insufficiently characterized.

LLM-driven Multi-agent Systems LLM-based multi-agent systems are increasingly used to support collective decision-making through communication, role differentiation, and coordination (Shen et al., 2023; Wang et al., 2025). Prior work suggests that interaction can trigger social-influence pitfalls in LLM groups, yielding majority amplification and erroneous consensus (Cui et al., 2025; Demszky et al., 2023). Yet most studies emphasize model performance and system design rather than operationalizing group-level constructs grounded in social and organizational psychology (Zhu et al., 2025). We bridge this gap by systematically evaluating groupthink and its typical associated phenomena in LLM collectives under controlled interaction settings.

Limitations and Future Work

GILCID is grounded in classical groupthink theory and experiments. However, classical groupthink theory and its canonical experimental paradigms are necessarily controlled abstractions and therefore cannot fully reflect the contextual richness of real-world groups (Janis, 2008). Future work should evaluate groupthink in more naturalistic collaborative settings and test additional factors emphasized in group psychology (such as group cohesion, cultural orientation, and member heterogeneity) and their relative contributions (Priyanto & Wening, 2024).

Conclusion

This work shows that groupthink and its canonical associated phenomena are prevalent in LLM-driven multi-agent decision-making and dialog. Using the GILCID framework and quantitative metrics, we demonstrate that these dynamics emerge systematically across models and tasks, positioning groupthink as a system-level social bias arising from interaction rather than an isolated model behavior. We further show that training-free group-level interaction interventions can mitigate these collective biases. Together, our findings underscore the need to study LLM-based multi-agent systems as computational social systems and provide a framework and metrics, grounded in groupthink theory, that serve as a reference point for measuring and moderating group-psychological dynamics in LLM collectives.

Acknowledgments

This work was supported by the Ministry of Public Security of China [2024ZB02].

References

Baltaji, R., Hemmatian, B., & Varshney, L. R. (2024). Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent llm collaboration. *The 2nd Workshop on Cross-Cultural Considerations in NLP*, 17.

Bernardelle, P., Civelli, S., Fröhling, L., et al. (2025). Political ideology shifts in large language models. https://arxiv.org/abs/2508.16013

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy*, *102*(5), 841–877.

Boone, T. (2005). Too much conformity leads to groupthink and failure. *Professionalization of Exercise Physiology*, *8*(9).

Conobi. (2018). PolitiScales - About — politiscales.party [[Accessed 01-10-2024]].

Cui, Z., Li, N., Zhou, H., et al. (2025). A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*.

Demszky, D., Yang, D., Yeager, D. S., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 14.

Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational behavior and human decision processes*, *73*(2-3), 116–141.

Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring massive multitask language understanding. *9th International Conference on Learning Representations*.

Janis, I. L. (2008). Groupthink. *IEEE Engineering Management Review*, *36*(1), 36.

Keren, A. (2025). Expert authority and its assessment. *Social Epistemology*, 1–14.

Liu, B., Jiang, Y., Zhang, X., et al. (2023). Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological bulletin*, *83*(4), 602.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43–51.

Priyanto, E., & Wening, N. (2024). The influence of dominant leadership and group cohesiveness on groupthink phenomenon in the decision-making process: A systematic literature review. *Eduvest-Journal of Universal Studies*, *4*(6), 4766–4784.

Rasal, S., & Hauer, E. (2024). Navigating complexity: Orchestrated problem solving with multi-agent llms. *arXiv preprint arXiv:2402.16713*.

Rose, J. D. (2011). Diverse perspectives on the groupthink theory–a literary review. *Emerging Leadership Journeys*, *4*(1), 37–57.

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *The Twelfth International Conference on Learning Representations*.

Shen, Y., Song, K., Tan, X., et al. (2023). Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, *36*, 38154–38180.

Suzgun, M., Scales, N., Schärli, N., et al. (2023). Challenging BIG-bench tasks and whether chain-of-thought can solve them. *Findings of the Association for Computational Linguistics: ACL 2023*, 13003–13051.

Talmor, A., Herzig, J., Lourie, N., et al. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4149–4158.

Wang, Q., Wang, T., Tang, Z., et al. (2025). Megaagent: A large-scale autonomous llm-based multi-agent system without predefined sops. *Findings of the Association for Computational Linguistics: ACL 2025*, 4998–5036.

Weng, Z., Chen, G., & Wang, W. (2025). Do as we do, not as you think: The conformity of large language models. *The Thirteenth International Conference on Learning Representations*.

Zhang, J., Xu, X., Zhang, N., et al. (2024). Exploring collaboration mechanisms for llm agents: A social psychology view. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14544–14607.

Zheng, M., Pei, J., & Jurgens, D. (2023). Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *CoRR*.

Zhu, X., Zhang, C., Stafford, T., et al. (2025). Conformity in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3854–3872.