SOCIALMAD: SOCIAL LEARNING-BASED MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Agent Debate (MAD) uses diverse independent judgments to provide complementary perspectives and correct reasoning errors. However, more interaction does not necessarily improve collaboration: group information can resolve errors but also interfere with agents that are already correct. Existing methods typically share all raw responses and use fixed rules such as conformity, which can interfere with valuable judgments and lead to erroneous consensus. Thus, the key to MAD is not more interaction, but enabling agents to learn how to use the resulting group information and turn diverse judgments into effective reasoning gains. From a social learning perspective, we propose SocialMAD, which reformulates collaboration in MAD as the optimization of personalized social learning strategies. SocialMAD organizes mixed raw responses into structured social information, encodes behaviors such as conformity and self-persistence as composable strategy units, and learns agent-specific social learning strategies from limited training data. These strategies guide agents to adaptively use group information during collaborative reasoning without predefined uniform rules. Experiments show that SocialMAD improves accuracy by up to 10.00% over the strongest baseline and increases the error-correction rate by approximately 19.38%. The learned strategies transfer to unseen tasks and benefit from more agents and interaction rounds.

1 INTRODUCTION

Large language models (LLMs) perform strongly on complex reasoning tasks (Wei et al., 2022), yet single trajectory generation makes local errors hard to correct (Wang et al., 2023; Madaan et al., 2023). Multi-Agent Debate (MAD) enables complementary reasoning and mutual correction by exchanging independent judgments and reasoning information among agents, and has emerged as an important multi-agent collaboration paradigm (Guo et al., 2024; Fan et al., 2026). However, more interaction does not always improve collaboration. Group information can help agents correct their judgments while exposing other correct agents to interference from erroneous opinions. Thus, the key to MAD is to transform diverse group judgments into effective reasoning gains for each agent, which depends on how group information is organized and how agents use it.

Early MAD methods (Smit et al., 2024; Chan et al., 2024) directly shared raw agent responses, risking redundancy and error propagation (Liang et al., 2024; Wang et al., 2024; Zhang et al., 2025). Later work optimizes information propagation through sparse communication (Li et al., 2024), topology learning (Sun et al., 2025) and dynamic routing (Yue et al., 2025), or filters redundant or unreliable information through response selection (Zeng et al., 2025), pruning (Zhang et al., 2025) and masking (Tian et al., 2026). However, they treat responses as interaction units, causing identical views to be repeatedly reinforced while conflicting views interfere with reasoning, potentially inducing erroneous consensus. Yet deduplication discards group-support cues. Therefore, organizing mixed responses into reliable group information is fundamental to effective utilization.

Existing studies have explored how agents utilize group information, typically guiding selection and updates with rules such as majority support (Choi et al., 2025) and high confidence (Chen et al., 2024; Zhu et al., 2026; Yoa et al., 2026). However, unified rules are often unsuitable for all agents. For example, majority rules may induce conformity, trapping incorrect agents in an erroneous majority while causing correct minority agents to abandon their judgments. Thus, shifting from unified information use rules to personalized strategies, enabling group interaction to promote

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

error correction while reducing interference with correct judgments, is a key challenge for MAD to translate group collaboration into improved individual capability.

Humans can extract valuable information from complex environments through social learning for effective problem solving (Morgan et al., 2011). Others' judgments and group support constitute social information available to individuals (Mesoudi, 2011). Social learning is not simply following the majority or consistently trusting a particular source (Behrens et al., 2008). Instead, individuals combine their judgments with social information and learn from feedback when to maintain their judgments or incorporate external opinions (Mahmoodi et al., 2018; Pereg et al., 2024). This motivates us to rethink collaboration in MAD: its core should not merely optimize group information interaction, but enable different agents to learn to use group information.

From a social learning perspective, we redefine multi-agent collaboration in MAD as the optimization of personalized social learning strategies and propose Social Learning-based Multi-Agent Debate (SocialMAD). SocialMAD constructs a structured representation of social information by grouping raw responses by answer consistency and integrating answers, key reasoning evidence, and group-support cues, shifting the collaboration unit from individual responses to competing answer clusters. This avoids repeated reinforcement of identical views, distinguishes conflicting views and transforms mixed group information into comparable and selectable social information.

Building on this social information, SocialMAD incorporates strategies, including self-persistence and conformity, into a unified strategy space. Each agent learns a lightweight personalized strategy offline using outcome supervision from limited training data. During inference, each agent combines its judgment with structured social information and uses the learned strategy to guide subsequent reasoning revision. The strategy learns to use social information rather than task-specific knowledge or answer patterns, enabling cross-task transfer with limited supervision. SocialMAD enables agents to adopt group consensus when reliable, preserve valuable minority opinions when the consensus is misleading, and avoid blind conformity when information conflicts. Our contributions are:

- **Rethinking MAD from a social learning perspective.** We shift MAD from how group information is exchanged to how individuals learn to use it, moving the optimization target from information interaction mechanisms to agents' social learning strategies and establishing a new modeling perspective for MAD.
- **Breaking fixed collaboration rules with social learning strategies.** SocialMAD organizes mixed group responses into structured social information, encodes behaviors such as conformity and self-persistence as composable strategy units, and learns personalized strategies from task feedback to guide agents in effectively using group information, reducing erroneous consensus while preserving valuable judgments.
- **Achieving substantial collaborative gains.** SocialMAD improves accuracy by up to 10.00% over the strongest baseline and raises the error-correction rate by up to 19.38% across diverse settings. The learned strategies transfer directly to unseen tasks and continue to benefit from more agents and interaction rounds.

2 RETHINKING MAD

2.1 PROBLEM AND OBJECTIVE FORMULATION

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces. A MAD system consists of N agents, $\mathcal{A} = \{a_1, \ldots, a_N\}$. Given an input $x \in \mathcal{X}$ with ground-truth answer $y^* \in \mathcal{Y}$, each agent a_i independently generates an initial response $m_i^0 = G_i(x) = \left(y_i^0, r_i^0, \rho_i^0\right)$, where G_i is the generation process of a_i, y_i^0 and r_i^0 are its answer and reasoning, and $\rho_i^0 \in [0,1]$ denotes the generation confidence of m_i^0.

At each round $t = 1, \ldots, T$, let $M^{t-1} = (m_1^{t-1}, \ldots, m_N^{t-1})$ denote the previous-round agent responses. Agent a_i then reasons over the problem x and information from M^{t-1} to update its response: $m_i^t = G_i\left(x, \mathcal{U}_i\left(M^{t-1}\right)\right), t = 1, \ldots, T$, where $\mathcal{U}_i$ characterizes how M^{t-1} is utilized by agent a_i. Conventional MAD directly uses all previous-round responses: $\mathcal{U}_i^{\text{raw}}(M^{t-1}) = M^{t-1}, \forall i$. For a shared rule F, all agents follow the same information-utilization rule. For example, conformity gives $\mathcal{U}_i^F(M^{t-1}) = M_{\text{maj}}^{t-1}, \forall i$, where M_{maj}^{t-1} denotes responses supporting the most-supported answer. After T rounds, the system produces the final group prediction $\widehat{y}^T = \text{Agg}\left(y_1^T, \ldots, y_N^T\right)$,

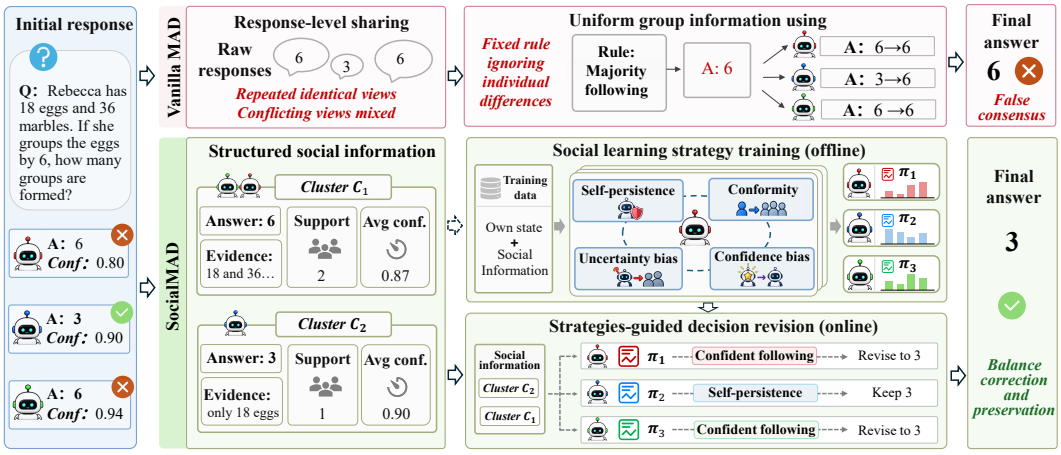


Figure 1: Framework comparison of Vanilla MAD and SocialMAD. Vanilla MAD directly shares raw responses or applies uniform information-utilization rules. SocialMAD organizes group responses into structured social information and learns personalized social learning strategies to guide agent-specific reasoning revision.

where $\mathrm{Agg} : \mathcal{Y}^N \to \mathcal{Y}$ denotes the answer aggregation function, typically majority voting. Regardless of how $\mathcal{U}_i$ is instantiated, MAD maximizes the expected task reward: $J_{\mathrm{MAD}} = \mathbb{E}\left[\mathcal{R}\left(\hat{y}^T, y^*\right)\right]$, where $\mathcal{R}(\cdot)$ denotes the task reward, e.g., answer correctness.

2.2 What MAD Lacks Is Appropriate Use of Group Information

The above formulation shows that utilization schemes $\mathcal{U}$ shape reasoning contexts and agent performance. Thus, MAD hinges on how group information is represented and used by different agents. We examine two typical schemes.

Observation 1: Sharing all raw responses conflates response count with independent viewpoints, amplifying repeated errors. Conventional MAD uses all previous-round responses, $\mathcal{U}_i^{\mathrm{raw}}(M^{t-1}) = M^{t-1}$. When multiple agents support the same incorrect answer, repeated presentation of that viewpoint and reasoning amplifies its influence on subsequent reasoning. As shown in Fig. 1, this causes the initially correct Agent 2 to switch from 3 to wrong answer 6. Fig. 2 shows that raw-response sharing turns 39.51% of initially correct judgments into incorrect ones. Thus, unorganized raw-response sharing can amplify common errors, suppress valuable divergence and drive the group toward erroneous consensus.

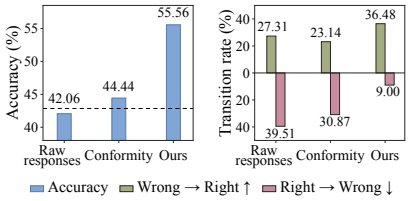


Figure 2: Comparison of SocialMAD with raw-response sharing and a fixed rule on MMLU-Logic. Left: accuracy. Right: wrong-to-right and right-to-wrong judgment transitions.

Observation 2: A uniform rule for utilizing information overlooks differences in agent states, making it difficult to balance error correction and correct-judgment preservation. When all agents follow the same rule, such as conformity, $\mathcal{U}_i^F(M^{t-1}) = M_{\mathrm{maj}}^{t-1}$, the same group information may not have the same reference value for agents in different states. Incorrect agents need external judgments for correction, whereas correct agents need to preserve their own judgments. As shown in Fig. 1, the conformity rule points to the incorrect answer 6, leaving incorrect agents uncorrected while causing the initially correct Agent 2 to abandon answer 3. Fig. 2 further shows that conformity corrects 23.14% of initially incorrect judgments but makes 30.87% of initially correct judgments incorrect. Thus, group information utilization strategies should be personalized to agent states.

2.3 Redefining MAD as Social Learning

These observations show that MAD hinges on group-information representation and agent-specific utilization. We reformulate MAD from a social learning perspective. Diverse group viewpoints and

support cues constitute social information that agents should use jointly with their own judgments. First, we organize the diverse candidate judgments and group-support cues from previous-round responses M^{t-1} into social information $\mathcal{G}^t$, replacing mixed raw responses with comparable candidate judgments and social cues. Meanwhile, $p_i^t = (y_i^{t-1}, \rho_i^{t-1})$ denotes the current state of agent a_i. We then model $\mathcal{U}_i$ as an agent-specific social learning strategy $\pi_i(\cdot \mid p_i^t, \mathcal{G}^t)$, rather than a predefined uniform rule, which determines how the agent uses candidate judgments to guide subsequent reasoning. Let $\Pi = \{\pi_1, \ldots, \pi_N\}$ denote the social learning strategies of all agents. The optimization objective is $\max_{\Pi} \; J_{\text{SL-MAD}}(\Pi) = \mathbb{E}\big[\mathcal{R}\big(\widehat{y}^T, y^*\big)\big]$. Accordingly, MAD is reformulated as a social learning process, shifting from predefined information use to personalized social learning strategies.

3 METHOD

SocialMAD has two components: structured social information representation and personalized social learning. As illustrated in Fig. 1, the latter involves offline strategy optimization and strategy-guided reasoning revision during inference. For simplicity, we describe a generic interaction round and omit round indices t on the current-round variables.

3.1 STRUCTURED SOCIAL INFORMATION REPRESENTATION

Social learning requires distinguishable social viewpoints and corresponding social evidence rather than unorganized individual responses. Vanilla MAD mixes repeated or conflicting viewpoints in the same context. We therefore recast M^{t-1} from response-level representations into viewpoint-level social information. Specifically, we extract and normalize answers, then group responses with semantically equivalent judgments into K mutually exclusive candidate clusters $\mathcal{C} = \{C_k\}_{k=1}^K, C_k = (y_k, e_k)$, where y_k is the candidate answer and e_k denotes its supporting reasoning evidence. For multiple rationales supporting the same answer, rather than directly exposing raw reasoning traces, we use $e_k = \text{Org}(\{r_i \mid y_i \sim_x y_k\})$ to organize the supporting evidence consistent with the candidate answer. Here, $y_i \sim_x y_k$ indicates that the two answers are semantically equivalent under the evaluation criterion of problem x. Thus, repeated expressions of the same viewpoint are grouped into the same candidate cluster while their valid supporting evidence is preserved.

Social information should capture both group viewpoints and their relative support. This is a key distinction between social learning and simple response compression: the same candidate answer can carry different social meaning under different support patterns. For candidate cluster C_k, let $\mathcal{N}_k = \{i \mid y_i \sim_x y_k\}$ denote its supporter set. We characterize its group support using the support size $n_k = |\mathcal{N}_k|$, mean generation confidence $\bar{\rho}_k = \frac{1}{n_k} \sum_{i \in \mathcal{N}_k} \rho_i$. Social learning focuses on the relative competition among candidate viewpoints, which isolated statistics cannot fully capture. We further construct two comparable social cues relative to the strongest competing candidate $\Delta n_k = \frac{n_k - \max_{l \neq k} n_l}{N}, \Delta\bar{\rho}_k = \bar{\rho}_k - \max_{l \neq k} \bar{\rho}_l$. When $K = 1$, both cues are set to zero. Accordingly, the social cues of candidate answer cluster C_k are represented as

$$\mathbf{g}_k = [\Delta n_k,\, \Delta\bar{\rho}_k]^\top , \tag{1}$$

which respectively characterize relative group support and relative group confidence. Finally, M^{t-1} is represented as structured social information:

$$\mathcal{G} = \{(C_k, \mathbf{g}_k)\}_{k=1}^K, \tag{2}$$

where each candidate cluster C_k captures a distinct viewpoint and its supporting reasoning evidence, while the social cues $\mathbf{g}_k$ characterize the relative support among these viewpoints, providing comparable social evidence for subsequent social learning strategies.

3.2 SOCIAL LEARNING STRATEGY OPTIMIZATION

Social learning suggests that individuals selectively use social information according to their own states rather than uniform rules. We therefore construct and optimize a personalized social learning strategy for each agent based on its state and social information to guide subsequent reasoning.

Social Learning Strategy Space Construction. Motivated by established social learning theory (Behrens et al., 2008), we instantiate four basic strategy units for MAD: self-persistence, copy-

when-uncertain, conformity and confidence-based following, forming a unified social learning strategy space, denoted as $\mathcal{S} = \{s_{\text{self}}, s_{\text{unc}}, s_{\text{con}}, s_{\text{conf}}\}$. These strategy units are not mutually exclusive rules, but composable information-utilization behaviors that characterize agents' reliance on individual and social information.

For agent a_i, each basic strategy's role depends on its state p_i and the social cues $\mathbf{g}_k$ of candidate C_k. Accordingly, we characterize its applicability to C_k using the strategy feature vector

$$\boldsymbol{\phi}_{i,k} = [\iota_{i,k}, (1-\rho_i)(1-\iota_{i,k}), \Delta n_k, \Delta\bar{\rho}_k]^\top , \tag{3}$$

where $\iota_{i,k} = \mathbb{I}[y_i \sim_x y_k]$ is the consistency between the candidate answer and agent a_i. Following the order of the strategy units in $\mathcal{S}$, the four components of $\boldsymbol{\phi}_{i,k}$ respectively characterize the applicability of each basic social learning strategy to candidate k for agent i in the current context.

Different agents should develop personalized social learning strategies. We parameterize the social learning strategy π_i of agent a_i as an agent-specific strategy weight vector $w_i \in \mathbb{R}^4$, where the four dimensions correspond one-to-one to the four basic strategy units in $\mathcal{S}$ and respectively represent agent a_i's personalized preferences associated with the corresponding basic strategy.

Since structured social information takes competing candidate answers as its basic units, the use of information by a social learning strategy is further reflected in its relative preference for different candidate answers in the current context. Accordingly, we weight the feature values of the basic strategies for candidate C_k by the agent's personalized strategy preferences, and define the strategy preference score of candidate C_k as

$$Q_{i,k} = w_i^\top \boldsymbol{\phi}_{i,k} \tag{4}$$

A larger $Q_{i,k}$ indicates that candidate C_k, under the current individual state and social information, is more consistent with agent a_i's personalized information-utilization preferences. Accordingly, the relative utilization tendency of agent a_i's personalized social learning strategy π_i toward candidate C_k is defined by normalizing the strategy preference scores:

$$\pi_i(k \mid p_i, G) = \frac{\exp\left(Q_{i,k}/\tau\right)}{\sum_{l=1}^{K} \exp\left(Q_{i,l}/\tau\right)}, \tag{5}$$

where $\tau > 0$ is the temperature. Thus, π_i is a preference distribution over candidate answers.

Personalized Strategy Optimization. These social learning preferences cannot be predefined, as agents differ in when to maintain their judgments and when to rely on group information. We therefore use limited data with outcome feedback to let each agent learn its social learning strategy π_i from actual task outcomes.

Given a small training set $\mathcal{D}$, agents independently generate initial responses M^0 and construct $\mathcal{G}$ for each sample (x, y^*). If a candidate C_{k^*} satisfies $y_{k^*} \sim_x y^*$, its index k^* is used as the supervision target. Otherwise, the sample is excluded. The optimization loss for strategy π_i is defined as $\ell^i_{\text{sel}}(x, y^*) = -\log \pi_i(k^* \mid p_i, \mathcal{G})$, which indirectly optimizes the agent's personalized strategies by increasing the preference probability of the correct candidate. Define $\mathcal{D}_{\text{sel}} = \{(x, y^*) \in \mathcal{D} | \exists k : y_k \sim_x y^*\}$ as the set of valid training samples and $W = (w_1, \ldots, w_N)$ as the strategy parameters of all agents. The overall strategy optimization objective is then defined as

$$\mathcal{L}_{\text{sel}}(W) = \frac{1}{N|\mathcal{D}_{\text{sel}}|} \sum_{i=1}^{N} \sum_{(x,y^*)\in\mathcal{D}_{\text{sel}}} \ell^i_{\text{sel}}(x, y^*). \tag{6}$$

Outcome supervision does not teach task knowledge or answer patterns. It identifies which information-utilization behaviors lead to valuable social information under a given individual state and social context. The strategy model takes only individual states and social cues as input, without encoding problem semantics or candidate-answer content, while the LLM remains frozen. Thus, optimization learns preferences for using social information rather than task-solving ability.

3.3 Strategy-Guided Collaborative Reasoning

After offline optimization, the learned strategy parameters w_i are frozen and reused across all interaction rounds on unseen problems to guide reasoning revision. For each interaction round, the current structured social information $\mathcal{G}$ is constructed from the previous-round responses M^{t-1}, and

the strategy feature $\phi_{i,k}$ is formed based on the agent's current state p_i and the social cues $\mathbf{g}_k$ of C_k. The agent a_i then forms preferences over the candidate answers according to its strategy π_i and selects the candidate with the highest strategy preference as the current social reference:

$$k_i = \arg\max_{1\le k\le K} \pi_i(k \mid p_i, \mathcal{G}) = \arg\max_{1\le k\le K} Q_{i,k}. \tag{7}$$

The selected C_{k_i}, containing the answer and consolidated evidence, serves as the social reference for revision. The learned strategy thus instantiates information utilization as $\mathcal{U}_i^{\pi}(M^{t-1}) = C_{k_i}$. Agent a_i then updates its response as $m_i^t = G_i(x, C_{k_i})$. If the learned strategy favors self-persistence under the current state and social information, it selects the cluster consistent with its current judgment to reinforce it. Otherwise, it selects an alternative cluster for reasoning revision. This repeats until round T. Overall, SocialMAD organizes group responses into structured social information and uses personalized strategies to guide agent-specific reasoning revision. This translates group interaction into stable individual reasoning gains and contributes to the task reward $J_{\text{SL-MAD}}$.

4 Theoretical Analysis

This section analyzes the theoretical advantages of personalized social learning. For any fixed rule F, let $\mathbf{e}_F \in \mathbb{R}^4$ denote the coordinate unit vector of its corresponding basic strategy unit. Define the fixed, shared and personalized strategy classes as $\mathcal{W}_F = \{(\alpha\mathbf{e}_F, \dots, \alpha\mathbf{e}_F) : \alpha > 0\}, \mathcal{W}_{\text{sh}} = \{(w, \dots, w) : w \in \mathbb{R}^4\}, \mathcal{W}_{\text{per}} = (\mathbb{R}^4)^N$. For agent i, define $L_i(w) = \mathbb{E}\big[\ell_i^{\text{sel}}(x, y^*; w)\big]$ as its population strategy loss over valid problems. For $W = (w_1, \dots, w_N)$, define $L(W) = \frac{1}{N}\sum_{i=1}^N L_i(w_i)$. Let L_F^*, L_{sh}^*, and L_{per}^* denote the infima of $L(W)$ over $\mathcal{W}_F$, $\mathcal{W}_{\text{sh}}$, and $\mathcal{W}_{\text{per}}$, respectively.

Proposition 1 (Fixed and shared strategies are degenerate cases of personalized strategies)**.** *The three strategy classes satisfy $\mathcal{W}_F \subseteq \mathcal{W}_{\text{sh}} \subseteq \mathcal{W}_{\text{per}}$, and consequently, $L_{\text{per}}^* \le L_{\text{sh}}^* \le L_F^*$.*

Theorem 2 (Lower Bound on the Gain from Personalized Strategies)**.** *Let $L_{\text{sh}}(w) = \frac{1}{N}\sum_{i=1}^N L_i(w)$. Suppose that the optimal population losses of the fixed rule F and the shared combination strategy are attained at finite parameters $w_F^* = \alpha^*\mathbf{e}_F$ and w_{sh}^*, respectively. Then $L_F^* - L_{\text{per}}^* \ge \frac{\tau^2}{8}\|\nabla L_{\text{sh}}(w_F^*)\|_2^2 + \frac{\tau^2}{8N}\sum_{i=1}^N \|\nabla L_i(w_{\text{sh}}^*)\|_2^2$. Here, $\nabla L_{\text{sh}}(w_F^*) \neq 0$ indicates that the fixed-rule loss can be reduced by combining other basic strategies. If there exists an agent i such that $\nabla L_i(w_{\text{sh}}^*) \neq 0$, the shared strategy cannot be optimal for all agents simultaneously, and personalization further reduces the loss.*

Theorem 3 (Finite-Sample Learning Guarantee for Personalized Strategies)**.** *Assume $\mathcal{D}_{\text{sel}}$ consists of i.i.d. samples of valid problems and the optimal personalized population loss L_{per}^* is attained by a finite parameter tuple $W_{\text{per}}^* = (w_{1,\text{per}}^*, \dots, w_{N,\text{per}}^*) \in \mathcal{W}_{\text{per}}$. For the cross-entropy strategy objective in Eq. (6), starting from a fixed initialization w_i^0 independent of the training samples, perform S steps of full-batch gradient descent on each agent's empirical loss with step size $\eta \le \tau^2/3$, and let $q = \eta S$. Denote the resulting strategy parameters by $\widehat{W} = (\widehat{w}_1, \dots, \widehat{w}_N)$. Then, $\mathbb{E}\Big[L(\widehat{W})\Big] - L_{\text{per}}^* \le \frac{B_{\text{per}}^2}{2q} + \frac{32q}{\tau^2|\mathcal{D}_{\text{sel}}|}$, where the expectation is taken over the i.i.d. draw of $\mathcal{D}_{\text{sel}}$ and $B_{\text{per}}^2 := \frac{1}{N}\sum_{i=1}^N \big\|w_{i,\text{per}}^* - w_i^0\big\|_2^2$.*

Remark. Theorem 3 yields three implications: **(i) Sample-efficient personalized strategy learning.** Optimizing the bound gives $\mathbb{E}\Big[L(\widehat{W})\Big] - L_{\text{per}}^* \le \frac{8B_{\text{per}}}{\tau\sqrt{|\mathcal{D}_{\text{sel}}|}} = O\big(|\mathcal{D}_{\text{sel}}|^{-1/2}\big)$. Achieving an ϵ-near-optimal population strategy loss requires $O(\epsilon^{-2})$ valid training problems, with no explicit dependence on LLM size. SocialMAD can learn near-optimal personalized social strategies from limited feedback. **(ii) Personalization gains under finite samples.** Let $\Delta := L_{\text{sh}}^* - L_{\text{per}}^* > 0$. Then $L_{\text{sh}}^* - \mathbb{E}\Big[L(\widehat{W})\Big] \ge \Delta - \frac{8B_{\text{per}}}{\tau\sqrt{|\mathcal{D}_{\text{sel}}|}}$. When $|\mathcal{D}_{\text{sel}}| > \frac{64B_{\text{per}}^2}{\tau^2\Delta^2}$, learned personalized strategy strictly outperforms shared optimum in expectation. Since the finite-sample error decays as $O(|\mathcal{D}_{\text{sel}}|^{-1/2})$, any personalization gain $\Delta > 0$ can be retained in expectation. When this error is at most $\Delta/2$, at least half of the gain is retained in expectation, which also holds over the optimal fixed strategy. **(iii) Theoretical basis for judgment preservation and error correction.** The incorrect-reference selection probability of social strategy is bounded by $\Pr(k_i \neq k^*) \le L_i(w_i)/\log 2$. Moreover, $\Pr(k_i \neq k^*) = \Pr(y_i^0 = y^*)\Pr(k_i \neq k^* \mid y_i^0 = y^*) + \Pr(y_i^0 \neq y^*)\Pr(k_i \neq k^* \mid y_i^0 \neq y^*)$, where

Table 1: Accuracy (%) across two experimental settings. CoT, CoT-SC report average accuracy of all agents. For LASE, we report the average accuracy across all possible leader choices.

Method	GSM8K	MATH -500	AIME24	AIME25	SVAMP	FOLIO-Wiki	MMLU-Logic	PrOntoQA
`Qwen2.5-1.5B / Qwen2.5-Math-1.5B / Qwen2.5-Coder-1.5B`								
CoT	56.05	38.07	2.22	0.00	66.90	47.83	42.85	52.40
CoT-SC	62.50	42.47	4.44	0.00	73.23	52.25	43.91	53.80
Majority Vote	67.32	46.00	3.33	0.00	78.50	51.52	41.27	54.80
MAD	67.25	39.80	0.00	0.00	81.70	53.48	42.06	52.80
RECONCILE	72.78	50.80	0.00	0.00	84.50	54.35	53.17	54.80
MAD-Conformist	70.96	43.00	3.33	0.00	81.20	53.91	49.20	54.60
MAD-M^2	77.79	59.40	3.33	0.00	90.40	**61.96**	50.00	55.60
Sparse MAD	71.49	47.80	6.67	0.00	83.00	56.30	45.24	55.80
MasRouter	77.26	53.20	10.00	6.67	87.70	54.13	41.27	55.60
LASE	60.83	43.47	1.11	2.22	72.00	49.49	45.24	54.00
SocialMAD	**82.26**	**66.80**	**13.33**	**10.00**	**93.60**	60.87	**55.56**	**60.40**
`DeepSeek-Math-7B / Qwen2.5-Math-7B / Llama-3.1-8B`								
CoT	82.81	46.27	4.44	2.22	83.43	57.68	48.14	70.60
CoT-SC	86.75	48.93	3.33	2.22	87.20	61.01	49.20	72.53
Majority Vote	86.88	47.80	6.67	0.00	86.70	61.96	53.17	84.60
MAD	91.96	67.20	0.00	0.00	90.40	65.43	55.56	89.20
RECONCILE	92.42	60.00	3.33	0.00	91.20	65.22	54.76	89.20
MAD-Conformist	90.75	56.80	3.33	3.33	90.80	65.00	54.76	87.20
MAD-M^2	94.69	69.40	10.00	3.33	90.70	62.17	59.52	90.80
Sparse MAD	93.18	66.80	3.33	3.33	90.70	65.65	55.56	91.40
MasRouter	87.41	44.60	0.00	3.33	88.40	63.48	54.76	82.80
LASE	83.50	49.93	4.44	3.33	84.50	60.87	50.79	72.67
SocialMAD	**95.45**	**75.80**	**13.33**	**13.33**	**93.90**	**72.83**	**59.52**	**94.20**

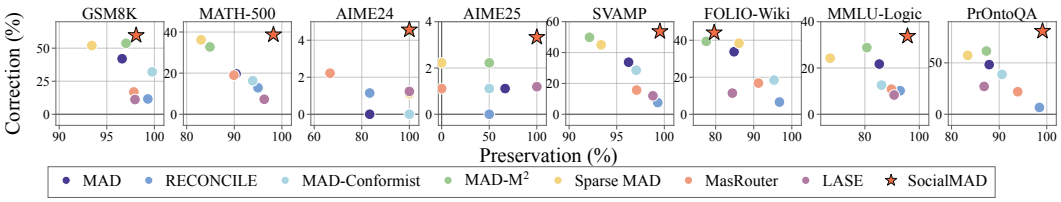


Figure 3: Agent-level judgment transitions under the cross-family setting. Points closer to the upper-right indicate better correct-judgment preservation and wrong-judgment correction.

the two terms capture state-weighted reference-selection errors, whose control supports correct-judgment preservation and error correction. Through $\mathcal{U}_i^{\pi}(M^{t-1}) = C_{k_i}$, learned personalized strategy utilizes social information to guide reasoning, helping improve collaborative objective $J_{\text{SL-MAD}}$. Full proofs are deferred to Appendix D.

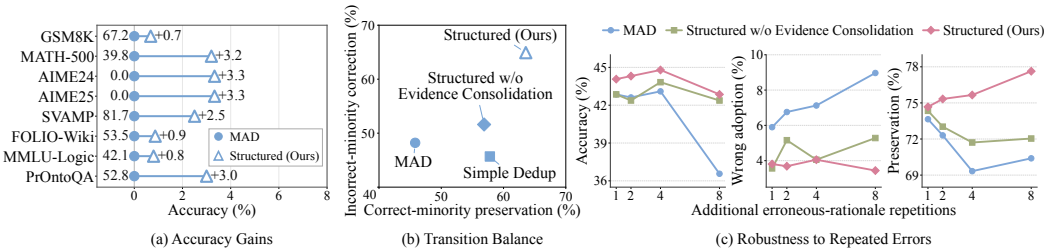
5 Experiments

5.1 Experimental Setup

Baselines. We compare SocialMAD with CoT (Wei et al., 2022), CoT-SC (Wang et al., 2023) as single-agent baselines. For multi-agent, we include Majority Vote, vanilla MAD (Du et al., 2024), RECONCILE (Chen et al., 2024), MAD-Conformist (Choi et al., 2025), MAD-M^2 (Tian et al., 2026), Sparse MAD (Li et al., 2024), MasRouter (Yue et al., 2025), LASE (Yoa et al., 2026).

Benchmarks. We evaluate on eight benchmarks spanning arithmetic reasoning (GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021)), competition math (MATH-500 (Hendrycks et al., 2021b; Lightman et al., 2024), AIME 2024/2025), logical reasoning (FOLIO-Wiki (Han et al., 2024; Wang et al., 2024), MMLU-Logic (Hendrycks et al., 2021a), PrOntoQA (Saparov & He, 2023)).

Implementation Details. We evaluate two heterogeneous agent settings: (i) same-family: Qwen2.5-1.5B, Qwen2.5-Math-1.5B, and Qwen2.5-Coder-1.5B (ii) cross-family: DeepSeek-Math-7B, Qwen2.5-Math-7B, and Llama-3.1-8B (Qwen et al., 2025; Yang et al., 2024; Hui et al., 2024; Shao et al., 2024; Grattafiori et al., 2024). All LLMs are frozen. Social strategies are trained on 100 samples from the GSM8K training set for the 1.5B setting and 100 samples from the MATH training set for the cross-family setting, which are then evaluated on all benchmarks. Details in Appendix E.



(a) Accuracy Gains (b) Transition Balance (c) Robustness to Repeated Errors

Figure 4: Comparison of social information representations: MAD uses raw responses; Simple Dedup removes duplicates; Structured w/o Evidence Consolidation clusters equivalent responses and Structured (Ours) further organizes supporting evidence. (a) Accuracy in the 1.5B setting. (b) Correct-minority preservation and incorrect-minority correction under majority–minority conflict. (c) Robustness to additional repeated erroneous rationales on MATH-500 in the cross-family setting.

5.2 Main Results

Overall Accuracy. As shown in Table 1, SocialMAD achieves the best accuracy in most settings. Specifically, SocialMAD outperforms the strongest baseline by 7.40% on MATH-500 in 1.5B setting and by 10.00% on AIME25 in cross-family setting. These results show SocialMAD improves overall collaborative performance, while the learned social learning strategies transfer effectively to unseen tasks, including logical reasoning tasks, demonstrating strong cross-task generalization.

Balance Between Error Correction and Correct-Judgment Preservation. As shown in Fig. 3, SocialMAD guides agents to effectively utilize group information through social strategies, achieving high error-correction and correct-judgment preservation rates across benchmarks. On PrOntoQA, SocialMAD achieves an 80.87% error-correction rate, improving over the strongest baseline MAD-M^2 by 19.38%, while preserving 99% of initially correct judgments. RECONCILE, LASE and MAD-Conformist apply uniform consensus- or confidence-based rules across agents, which are less adaptive to agent-specific states and limit error correction. MAD-M^2 and Sparse MAD filter information for revision but cannot determine whether it is corrective or misleading for the current agent, making correct judgments more vulnerable. MAD directly shares all raw responses and is therefore susceptible to interference from erroneous information.

5.3 Structured Social Information Mitigates Conflict Interference and Error Amplification

Structured Social Information Mitigates Interference from Raw Responses. As shown in Fig. 4(a), Structured social information improves accuracy over MAD across all benchmarks by up to 3.3%. In Fig. 4(b), Simple Dedup achieves 57.80% preservation but lower correction, suggesting deduplication discards useful reasoning information. Structured achieves 63.57% preservation and 64.92% correction. By clustering semantically equivalent responses while separating conflicting views and organizing their evidence, structured social information facilitates error correction and protects valuable minority judgments.

Robustness to Repeated Erroneous Rationales. To evaluate robustness against repetition-induced error amplification, we additionally repeat the same erroneous rationale 1-8 times in the social context. As shown in Fig. 4(c), as repetition increases, MAD accuracy drops to 36.56%, showing repeatedly presenting the same error increasingly amplifies its influence on subsequent reasoning. In contrast, Structured w/o Evidence Consolidation markedly reduces the degradation and Structured further achieves 42.86% accuracy after 8 repetitions, with only 3.44% wrong-answer adoption. These results show that semantic clustering prevents repeated errors from being treated as multiple independent viewpoints, while evidence consolidation further refines within-cluster evidence, jointly improving robustness to repeated erroneous rationales.

5.4 Social Learning Improves Answer Selection and Reasoning Correction

Personalized Strategies Outperform Fixed Rules. As shown in Fig. 5(a), SocialMAD achieves 93.35% correct-reference selection accuracy, outperforming Shared Policy and the best fixed strategy by 3.67% and 5.57%. It also yields the highest final accuracy 94.20%, showing that improved

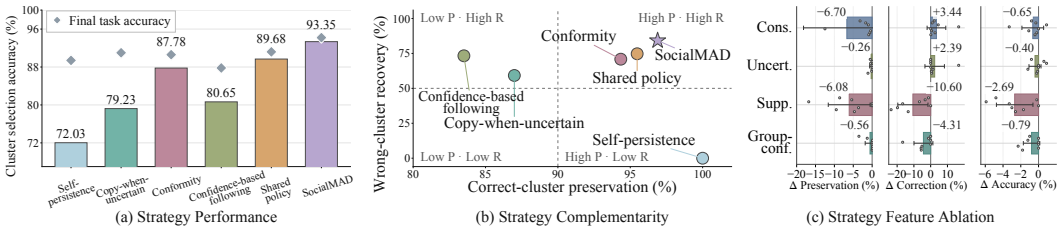


Figure 5: Comparison of fixed, shared and personalized social learning strategies. (a) Correct-cluster selection accuracy (bars), overall accuracy (diamonds) on PrOntoQA under the cross-family setting. (b) Correct-cluster preservation and wrong-cluster recovery. (c) Strategy feature ablation. Each variant is retrained after removing one feature. Dots show per-benchmark performance changes relative to the full strategy, while bars denote the mean changes, with error bars indicating the standard deviation across benchmarks.

social-information utilization translates into better reasoning performance. Shared Policy combines multiple social learning behaviors but cannot capture agent-specific differences, whereas SocialMAD learns personalized strategies for more reliable reasoning revision.

Different Social Learning Strategies Provide Complementary Benefits. As shown in Fig. 5(b), fixed strategies exhibit complementary behaviors. Self-persistence favors preservation but offers little correction, while confidence-based following improves correction at the cost of preservation. Conformity sacrifices preservation, as consensus may mislead agents away from correct judgments. This trade-off indicates that no single social learning strategy suits all agent states. As individual states and social cues are encoded as features for each social learning strategy unit, Fig. 5(c) further examines the effects of removing each feature. On average, removing Consistency reduces correct-answer preservation by 6.70%, while removing Support reduces wrong-answer correction by 10.60% and final accuracy by 2.69%. Individual-state cues mainly support the preservation of existing judgments, whereas social cues provide external evidence for correction. Accordingly, the personalized social strategy combines the roles of these two types of cues on different social learning strategy units in the current context, thereby forming an information-utilization behavior adapted to the current state.

5.5 Scaling Analysis of Agent Number and Interaction Rounds

As shown in Fig. 6, SocialMAD benefits from scaling and accuracy rises from 75.80% to 79.20% as agents increase, while MAD falls from 67.20% to 60.00%. SocialMAD shows an overall upward trend across rounds, while MAD fluctuates. By structuring social information and adaptively utilizing it through social learning strategies, SocialMAD converts increased interaction into useful reasoning evidence, thereby improving performance.

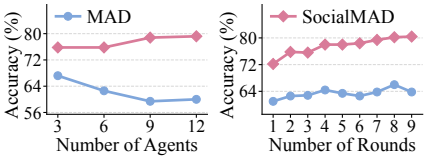


Figure 6: Scaling on MATH-500 under cross-family setting. Left: agents over 2 rounds. Right: rounds with 6 agents.

6 Conclusion

This paper revisits MAD from a social learning perspective, modeling multi-agent interaction as a process of social learning among agents. SocialMAD organizes raw responses into structured social information and combines social learning behaviors, such as self-persistence and conformity, into learnable personalized strategies for adaptive group information use. Theoretical analysis shows that personalized social learning subsumes fixed rules in the strategy space, provides improvement guarantees and admits a finite-sample learning guarantee. Experiments show that SocialMAD corrects erroneous judgments while preserving correct ones, transfers learned strategies to unseen tasks, and continues to yield collaborative gains as the number of agents and interaction rounds increase. Beyond improving a particular MAD method, SocialMAD introduces social learning into multi-agent reasoning, providing a new modeling framework for moving from groups of agents that interact to groups of agents that learn from one another.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and for retrieval and discovery. Specifically, all manuscript content was first written by the authors, and generative AI was used only to improve language and clarity. We used generative AI to search for and identify potentially relevant literature for the authors to review further. We did not use generative AI tools for research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims. No other uses requiring disclosure are applicable to this work. We reviewed and verified all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, citations, and artifacts.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce our results throughout the paper and appendix. The experimental setup, including the agent configurations, datasets, strategy training procedure, evaluation protocol, evaluation metrics, and implementation details, is described in Section 5.1 and Appendix E. Additional details of the proposed method are provided in Appendix B, and additional experimental results are reported in Appendix F. For the theoretical results, all assumptions and complete proofs are included in Appendix D.

REFERENCES

Timothy EJ Behrens, Laurence T Hunt, Mark W Woolrich, and Matthew FS Rushworth. Associative learning of social value. *Nature*, 456(7219):245–249, 2008.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *Proceedings of the International Conference on Learning Representations*, pp. 9079–9093, 2024.

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

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, pp. 101732–101764, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. 2021.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the International Conference on Machine Learning*, pp. 11733–11763, 2024.

Wei Fan, JinYi Yoon, and Bo Ji. iMAD: Intelligent multi-agent debate for efficient and accurate LLM Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35): 29403–29411, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. 2024.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 8048–8057, 2024.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm

Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, 2024.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the International Conference on Machine Learning*, volume 48, pp. 1225–1234, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*, pp. 1–27, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, pp. 1–11, 2021b.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. Multiple LLM agents debate for equitable cultural alignment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24841–24877, 2025.

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

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023.

Ali Mahmoodi, Bahador Bahrami, and Carsten Mehring. Reciprocity of social influence. *Nature Communications*, 9(1):2474, 2018.

Abdine Maiga, Anoop Shah, and Emine Yilmaz. Error detection in medical note through multi agent debate. In *Proceedings of the Workshop on Biomedical Language Processing*, pp. 124–135, 2025.

Alex Mesoudi. An experimental comparison of human social learning strategies: payoff-biased social learning is adaptive but underused. *Evolution and Human Behavior*, 32(5):334–342, 2011.

T. J. H. Morgan, L. E. Rendell, M. Ehn, W. Hoppitt, and K. N. Laland. The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, 279(1729): 653–662, 2011.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, 2021.

Maayan Pereg, Uri Hertz, Ido Ben-Artzi, and Nitzan Shahar. Disentangling the contribution of individual and social learning processes in human advice-taking behavior. *npj Science of Learning*, 9(1):4, 2024.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. 2025.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *Proceedings of the International Conference on Learning Representations*, pp. 1–20, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. 2024.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 45883–45905, 2024.

Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. CortexDebate: Debating sparsely and equally for multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9503–9523, 2025.

Hongduan Tian, Xiao Feng, Ziyuan Zhao, Xiangyu Zhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. In *Proceedings of the International Conference on Learning Representations*, pp. 5122–5143, 2026.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6106–6131, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations*, pp. 1–15, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Seungdong Yoa, Ye Seul Sim, Suhee Yoon, Sanghyu Yoon, Dongmin Kim, Soonyoung Lee, Bumsoo Kim, and Junhyun Lee. Efficient multi-agent reasoning via confidence-guided adaptive debate. In *Proceedings of the International Conference on Machine Learning*, pp. 1–18, 2026.

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. MasRouter: Learning to route LLMs for multi-agent systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15549–15572, 2025.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S^2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9393–9408, 2025.

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

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

A Related Works

A.1 Multi-Agent Debate

Multi-Agent Debate (MAD) enables multiple LLMs to independently form judgments and exchange reasoning information across subsequent rounds, allowing diverse reasoning trajectories to complement one another and provide external signals for correcting errors made by individual models (Du et al., 2024). As a representative paradigm for multi-agent collaborative reasoning, MAD has been applied to tasks such as model evaluation (Chan et al., 2024), medical reasoning (Maiga et al., 2025), and cross-cultural decision making (Ki et al., 2025).

Information organization and communication in MAD. Simply increasing interaction does not necessarily improve reasoning, as redundant communication and unfiltered information may instead reduce viewpoint diversity and propagate errors (Smit et al., 2024; Liang et al., 2024; Wang et al., 2024; Zhang et al., 2025). To address this issue, one line of work reduces unnecessary information exchange by optimizing when agents interact and how communication is structured. iMAD dynamically determines whether to initiate debate for a given problem (Fan et al., 2026). Sparse MAD limits inter-agent connectivity through sparse communication topologies (Li et al., 2024). CortexDebate dynamically constructs sparse debate graphs based on the potential helpfulness between agents (Sun et al., 2025). MasRouter jointly determines collaboration modes, role allocation, and LLM routing (Yue et al., 2025). Another line of work directly filters or compresses the exchanged content to reduce the propagation of redundant and low-value information. S^2-MAD reduces ineffective information exchange through interaction sparsification (Zeng et al., 2025). AgentPrune prunes redundant or potentially harmful communication messages (Zhang et al., 2025). MAD-M^2 employs memory masking to filter erroneous information from prior interactions (Tian et al., 2026). These methods improve the transmission, filtering, and interaction structure of group information, but typically still treat each individual response as the basic processing unit. Consequently, semantically identical judgments may be repeatedly presented as multiple independent responses, causing repetition itself to be interpreted as stronger independent support, thereby amplifying erroneous viewpoints and potentially driving the group toward erroneous consensus. Meanwhile, different or even conflicting viewpoints and their rationales remain mixed within the same context, making it difficult for agents to distinguish useful evidence from distracting information, which may both hinder the correction of erroneous judgments and mislead initially correct ones.

Group Information Utilization Mechanisms in MAD. Traditional MAD typically places the complete responses of all agents directly into the context, leaving the LLM to implicitly determine how to integrate this information (Du et al., 2024). However, collaborative performance depends not only on what information an agent receives, but also on which cues it relies on to utilize that information and revise its own judgment. Accordingly, subsequent studies have begun to explicitly design mechanisms for group-information utilization. ReConcile incorporates confidence information into multi-model discussions and promotes consensus formation through a confidence-weighted mechanism (Chen et al., 2024). MAD-Conformist uses majority support to impose a conformity bias on agents' judgment updates (Choi et al., 2025). Demystifying MAD further introduces confidence-modulated debate, allowing agents to adjust their adoption of external viewpoints according to the confidence of other agents (Zhu et al., 2026). LASE adaptively determines whether further debate is needed based on confidence states (Yoa et al., 2026). Cues such as majority support, confidence, and information source provide useful signals for agents to utilize group information. However, these methods typically adopt uniform information utilization rules and pay limited attention to the fact that different agents may require different information-utilization strategies. For example, conformity or confident-source following may help some incorrect agents revise their judgments using

group information. Whereas self-persistence may be more important for agents that have already formed correct judgments, conformity may instead cause such agents to abandon their correct judgments and follow an erroneous consensus. Therefore, a single information utilization strategy is unlikely to suit all agents. How to develop appropriate strategies for different agents, such that group information can facilitate error correction while allowing agents to preserve their own judgments when they are correct, remains an insufficiently addressed problem in MAD.

A.2 Social Learning

Social learning has studied how individuals acquire information from others' behaviors, judgments, and experiences and use it to adjust their own decisions, providing important insights into how different individuals can develop appropriate information-utilization strategies. Social learning does not simply amount to following the majority. Rather, individuals adopt different ways of utilizing information depending on their own states and available social cues. Behavioral studies have shown that reliance on social information is influenced by factors such as individual uncertainty, the degree of group consensus, source performance, and the task environment (Morgan et al., 2011). Mesoudi (2011) compared different social learning strategies, including payoff-biased learning and conformity, showing that different social cues can induce different patterns of information utilization. Behrens et al. (2008) further showed that individuals can learn the value of different social information sources from interaction feedback and adjust their subsequent decisions accordingly. Subsequent studies have further revealed the dynamic nature and individual differences in social-information utilization. Mahmoodi et al. (2018) showed that individuals' reliance on others' opinions changes with their interaction history. Pereg et al. (2024) found that people update their reliance on external advice based on past feedback, while exhibiting different advice-following tendencies across individuals. Together, these studies suggest that effective social learning requires jointly considering an individual's own state, social cues within the group, and individualized tendencies in information utilization. This provides MAD with a perspective beyond uniform interaction rules: group responses can be organized as social information, while different agents should adopt social learning strategies conditioned on their own states and social information.

B Structured Social Information Construction Details

B.1 Candidate Answer Clustering based on Semantic Equivalence

The clustering of candidate answers is determined by the semantic equivalence relation $y_i \sim_x y_j$, which is implemented as follows. For each agent response $m_i = (y_i, r_i, \rho_i)$, we first extract the final answer y_i and apply answer normalization. Specifically, we define a normalization function $\mathrm{Norm}(\cdot)$ to remove superficial formatting variations across different answer representations, such as whitespace and redundant symbols.

For numerical reasoning tasks, the normalized answers are further parsed into standard mathematical expressions:

$$E(y_i) = \mathrm{Parse}(\mathrm{Norm}(y_i)),$$

where $\mathrm{Parse}(\cdot)$ denotes the process of converting a normalized answer string into a mathematical expression. Two answers are considered semantically equivalent when:

$$y_i \sim_x y_j \iff E(y_i) = E(y_j).$$

For example, $E(0.5) = E(\frac{1}{2}) = E(1/2)$, and thus these different surface forms are grouped into the same candidate answer cluster.

For non-numerical reasoning tasks, where answers are typically represented as discrete labels or short texts, we directly determine semantic equivalence based on normalized answers:

$$y_i \sim_x y_j \iff \mathrm{Norm}(y_i) = \mathrm{Norm}(y_j).$$

B.2 Evidence Consolidation within Candidate Answer Clusters

For each candidate answer cluster C_k, we organize the reasoning associated with its supporting responses as $e_k = \mathrm{Org}\big(\{r_i \mid y_i \sim_x y_k\}\big)$, where $\mathrm{Org}(\cdot)$ denotes an evidence organization procedure

within the candidate cluster. Rather than directly passing the original reasoning traces as unstructured response-level information, the procedure organizes their supporting content with respect to the candidate answer y_k.

We implement $\mathrm{Org}(\cdot)$ using lightweight rule-based processing without additional LLM inference. Each reasoning trace is decomposed into local reasoning units based on its textual structure. We then filter segments that are inconsistent with the candidate answer or correspond to subsequently rejected reasoning, while retaining the remaining candidate-consistent content as supporting evidence.

The resulting reasoning units constitute the evidence e_k associated with candidate cluster C_k. When filtering would remove excessive context, we conservatively retain the original reasoning trace to avoid substantial information loss.

C Overall Algorithm

Algorithm 1 Offline Personalized Social Strategy Optimization

Require: Training set $\mathcal{D}$, agents $\mathcal{A} = \{a_1, \ldots, a_N\}$
Ensure: Personalized social strategy parameters $W = \{w_i\}_{i=1}^N$

```
Initialize valid training set D_sel ← ∅
for each training sample (x, y*) ∈ D do
    Each agent i ∈ {1, ..., N} independently generates m_i^0 = (y_i^0, r_i^0, ρ_i^0)
    Construct structured social information G = {(C_k, g_k)}_{k=1}^K
    if ∃ k such that y_k ∼_x y* then
        Identify k* such that y_{k*} ∼_x y*
        Add (x, y*) to D_sel
        Cache G, k*, and {φ_{i,k}}_{i,k}
    end if
end for
Initialize strategy parameters W
for each (x, y*) ∈ D_sel do
    for each agent i ∈ {1, ..., N} do
        Compute strategy scores {Q_{i,k} = w_i^⊤ φ_{i,k}}_{k=1}^K
        Compute strategy loss ℓ_i^sel = − log π_i(k* | p_i, G)
    end for
end for
Optimize W by minimizing L_sel(W) as defined in Eq. (6)
return W
```

Algorithm 2 Strategy-guided Collaborative Reasoning

Require: Problem x, agents $\mathcal{A} = \{a_1, \ldots, a_N\}$, learned social strategies $\{w_i\}_{i=1}^N$, rounds T
Ensure: Final prediction $\widehat{y}^T$

```
Each agent generates initial response m_i^0 = (y_i^0, r_i^0, ρ_i^0)
for t = 1, ..., T do
    Construct structured social information G^t from {m_i^{t−1}}_{i=1}^N
    for agent i = 1, ..., N do
        Compute candidate preference: π_i(k|p_i^t, G^t)
        Select social reference: k_i = arg max_k π_i(k|p_i^t, G^t)
        Revise response: m_i^t = G_i(x, C_{k_i})
    end for
end for
Aggregate final answers: ŷ^T = Agg(y_1^T, ..., y_N^T)
return ŷ^T
```

D PROOF OF SECTION 4

D.1 PROOF OF PROPOSITION 1

Proof. Since $\mathcal{W}_F = \{(\alpha \mathbf{e}_F, \dots, \alpha \mathbf{e}_F) : \alpha > 0\}, \mathcal{W}_{\text{sh}} = \left\{(w, \dots, w) : w \in \mathbb{R}^4\right\}, \mathcal{W}_{\text{per}} = (\mathbb{R}^4)^N$. For any $\alpha > 0$, the vector $\alpha \mathbf{e}_F$ belongs to $\mathbb{R}^4$. Therefore, $(\alpha \mathbf{e}_F, \dots, \alpha \mathbf{e}_F)$ is an element of $\mathcal{W}_{\text{sh}}$, and hence $\mathcal{W}_F \subseteq \mathcal{W}_{\text{sh}}$.

Furthermore, any shared parameter tuple $(w, \dots, w) \in \mathcal{W}_{\text{sh}}$ consists of N vectors in $\mathbb{R}^4$, and therefore also belongs to the Cartesian product $(\mathbb{R}^4)^N = \mathcal{W}_{\text{per}}$. Hence, $\mathcal{W}_{\text{sh}} \subseteq \mathcal{W}_{\text{per}}$.

We next translate these set-inclusion relations into inequalities between the corresponding losses. For any real-valued objective function L, if $\mathcal{V}_1 \subseteq \mathcal{V}_2$, then taking the infimum over the larger set cannot yield a larger value, i.e., $\inf_{W \in \mathcal{V}_2} L(W) \leq \inf_{W \in \mathcal{V}_1} L(W)$. Applying it first to $\mathcal{W}_{\text{sh}} \subseteq \mathcal{W}_{\text{per}}$ and then to $\mathcal{W}_F \subseteq \mathcal{W}_{\text{sh}}$, we obtain $\inf_{W \in \mathcal{W}_{\text{per}}} L(W) \leq \inf_{W \in \mathcal{W}_{\text{sh}}} L(W) \leq \inf_{W \in \mathcal{W}_F} L(W)$. By the definitions of L^*_{per}, L^*_{sh}, and L^*_F, the above inequality is exactly $L^*_{\text{per}} \leq L^*_{\text{sh}} \leq L^*_F$ □

D.2 PROOF OF THEOREM 2

Lemma 4 (Uniform boundedness of the strategy feature vectors)**.** *For any valid problem, any agent i, and any candidate C_k, $\|\boldsymbol{\phi}_{i,k}\|_2^2 \leq 3, \|\boldsymbol{\phi}_{i,k}\|_2 \leq \sqrt{3}$.*

Proof. By Eq. (3), $\boldsymbol{\phi}_{i,k} = [\iota_{i,k}, (1-\rho_i)(1-\iota_{i,k}), \Delta n_k, \Delta \bar{\rho}_k]^\top$. We bound the four components directly according to their definitions in the main text. First, $\iota_{i,k} = \mathbb{I}[y_i \sim_x y_k] \in \{0, 1\}, \rho_i \in [0, 1]$. The first two components cannot be nonzero simultaneously. Specifically, when $\iota_{i,k} = 1$, $(1-\rho_i)(1-\iota_{i,k}) = 0$; when $\iota_{i,k} = 0$, the first component is zero, while the second component is $1 - \rho_i \in [0, 1]$. Therefore,

$$\iota_{i,k}^2 + [(1-\rho_i)(1-\iota_{i,k})]^2 \leq 1. \tag{8}$$

Second, Eq. (1) defines $\Delta n_k = \frac{n_k - \max_{l \neq k} n_l}{N}$. Since each support count lies in $[0, N]$, we have $\Delta n_k \in [-1, 1]$. Similarly, $\bar{\rho}_k \in [0, 1]$, and hence its difference relative to the strongest competing candidate satisfies

$$\Delta \bar{\rho}_k \in [-1, 1]. \tag{9}$$

When $K = 1$, the main text sets both social cues to zero, so the above bounds still hold. Combining Eq. (8) and (9), we obtain $\|\boldsymbol{\phi}_{i,k}\|_2^2 \leq 1 + 1 + 1 = 3$, and therefore $\|\boldsymbol{\phi}_{i,k}\|_2 \leq \sqrt{3}$.

□

Lemma 5 (Regularity of the strategy loss)**.** *For any agent i and any valid problem, the single-sample strategy loss $\ell_i^{\text{sel}}(x, y^*; w)$ is convex, $4/\tau$-Lipschitz, and $3/\tau^2$-smooth with respect to $w \in \mathbb{R}^4$. Consequently, the empirical loss $\widehat{L}_i(w) = \frac{1}{|\mathcal{D}_{\text{sel}}|} \sum_{(x,y^*) \in \mathcal{D}_{\text{sel}}} \ell_i^{\text{sel}}(x, y^*; w)$ and the population loss $L_i(w)$ are also convex, $4/\tau$-Lipschitz, and $3/\tau^2$-smooth. Moreover, $L_{\text{sh}}(w) = \frac{1}{N} \sum_{i=1}^N L_i(w)$ is convex and $3/\tau^2$-smooth.*

Proof. Fix a valid problem and an agent i. Since the strategy features $\boldsymbol{\phi}_{i,k}$ are constructed before optimizing w, they are treated as fixed with respect to w. We omit (x, y^*) when no ambiguity arises. By Eqs. (4)–(5) and the definition of ℓ_i^{sel}, the single sample strategy loss can be written as

$$\ell_i^{\text{sel}}(w) = \log\left(\sum_{k=1}^K \exp\left(\frac{w^\top \boldsymbol{\phi}_{i,k}}{\tau}\right)\right) - \frac{w^\top \boldsymbol{\phi}_{i,k^*}}{\tau}. \tag{10}$$

For notational simplicity, we write $\pi_i(k \mid w) := \pi_i(k \mid p_i, \mathcal{G}; w)$ throughout this proof, with $(p_i, \mathcal{G})$ fixed. Differentiating Eq. (10) gives

$$\nabla \ell_i^{\text{sel}}(w) = \frac{1}{\tau}\left(\sum_{k=1}^K \pi_i(k \mid w) \boldsymbol{\phi}_{i,k} - \boldsymbol{\phi}_{i,k^*}\right). \tag{11}$$

By Lemma 4, $\|\boldsymbol{\phi}_{i,k}\|_2 \le \sqrt{3}$ for every candidate k. Since $\pi_i(k \mid w) \ge 0$ and $\sum_{k=1}^{K} \pi_i(k \mid w) = 1$, Eq. (11) yields

$$\left\|\nabla \ell_i^{\mathrm{sel}}(w)\right\|_2 \le \frac{1}{\tau}\left(\sum_{k=1}^{K} \pi_i(k \mid w)\|\boldsymbol{\phi}_{i,k}\|_2 + \|\boldsymbol{\phi}_{i,k^*}\|_2\right) \le \frac{2\sqrt{3}}{\tau} \le \frac{4}{\tau}. \tag{12}$$

Hence, by the mean-value inequality, ℓ_i^{sel} is $4/\tau$-Lipschitz.

We next establish convexity and smoothness. Differentiating the softmax probability gives

$$\nabla \pi_i(k \mid w) = \frac{\pi_i(k \mid w)}{\tau}\left(\boldsymbol{\phi}_{i,k} - \sum_{l=1}^{K} \pi_i(l \mid w)\boldsymbol{\phi}_{i,l}\right). \tag{13}$$

Substituting Eq. (13) into the derivative of Eq. (11) yields

$$\nabla^2 \ell_i^{\mathrm{sel}}(w) = \frac{1}{\tau^2}\left[\sum_{k=1}^{K} \pi_i(k \mid w)\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^{\top} - \left(\sum_{k=1}^{K} \pi_i(k \mid w)\boldsymbol{\phi}_{i,k}\right)\left(\sum_{k=1}^{K} \pi_i(k \mid w)\boldsymbol{\phi}_{i,k}\right)^{\top}\right]. \tag{14}$$

The matrix in brackets is the covariance matrix of the random vector that takes value $\boldsymbol{\phi}_{i,k}$ with probability $\pi_i(k \mid w)$. It is therefore positive semidefinite, which implies $\nabla^2 \ell_i^{\mathrm{sel}}(w) \succeq 0$. Hence, ℓ_i^{sel} is convex.

To bound the Hessian, let $v \in \mathbb{R}^4$ satisfy $\|v\|_2 = 1$. From Eq. (14),

$$v^{\top}\nabla^2 \ell_i^{\mathrm{sel}}(w)v = \frac{1}{\tau^2}\operatorname{Var}_{k\sim\pi_i(\cdot\mid w)}\left(v^{\top}\boldsymbol{\phi}_{i,k}\right). \tag{15}$$

Using $\operatorname{Var}(Z) \le \mathbb{E}[Z^2]$ and the Cauchy–Schwarz inequality,

$$\begin{aligned} v^{\top}\nabla^2 \ell_i^{\mathrm{sel}}(w)v &\le \frac{1}{\tau^2}\mathbb{E}_{k\sim\pi_i(\cdot\mid w)}\left[\left(v^{\top}\boldsymbol{\phi}_{i,k}\right)^2\right] \\ &\le \frac{1}{\tau^2}\mathbb{E}_{k\sim\pi_i(\cdot\mid w)}\left[\|v\|_2^2\|\boldsymbol{\phi}_{i,k}\|_2^2\right] \\ &\le \frac{3}{\tau^2}, \end{aligned} \tag{16}$$

where the last inequality follows from $\|v\|_2 = 1$ and Lemma 4.

Since the Hessian is positive semidefinite, the Rayleigh–Ritz characterization together with Eq. (16) gives $\left\|\nabla^2 \ell_i^{\mathrm{sel}}(w)\right\|_{\mathrm{op}} \le \frac{3}{\tau^2}$. Therefore, ℓ_i^{sel} is $3/\tau^2$-smooth.

Convexity, Lipschitzness, and smoothness are preserved under finite averaging, $\widehat{L}_i$ is also convex, $4/\tau$-Lipschitz, and $3/\tau^2$-smooth.

For the population loss, $L_i(w) = \mathbb{E}\left[\ell_i^{\mathrm{sel}}(x, y^*; w)\right]$. Eq. (12) provides a uniform bound on the single-sample gradient, so dominated convergence permits differentiation under the expectation:

$$\nabla L_i(w) = \mathbb{E}\left[\nabla \ell_i^{\mathrm{sel}}(x, y^*; w)\right].$$

For any $w, w' \in \mathbb{R}^4$, Jensen's inequality and the single-sample smoothness bound give

$$\begin{aligned} \|\nabla L_i(w) - \nabla L_i(w')\|_2 &\le \mathbb{E}\left[\left\|\nabla \ell_i^{\mathrm{sel}}(w) - \nabla \ell_i^{\mathrm{sel}}(w')\right\|_2\right] \\ &\le \frac{3}{\tau^2}\|w - w'\|_2. \end{aligned} \tag{17}$$

Thus, L_i is $3/\tau^2$-smooth. Taking expectations also preserves convexity and the $4/\tau$-Lipschitz bound.

Finally, $\nabla L_{\mathrm{sh}}(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla L_i(w)$, and therefore

$$\begin{aligned} \|\nabla L_{\mathrm{sh}}(w) - \nabla L_{\mathrm{sh}}(w')\|_2 &\le \frac{1}{N}\sum_{i=1}^{N} \|\nabla L_i(w) - \nabla L_i(w')\|_2 \\ &\le \frac{3}{\tau^2}\|w - w'\|_2. \end{aligned} \tag{18}$$

Hence, L_{sh} is also convex and $3/\tau^2$-smooth. □

Proof. The proof of Theorem 2 proceeds in three steps. We first establish a descent bound for a general smooth function, then apply it to the two relaxations from a fixed rule to a shared combination strategy and from a shared combination strategy to personalized strategies, and finally combine the two gains.

If a differentiable function f has an L_s-Lipschitz gradient, then for any u, v,

$$f(v) \le f(u) + \langle \nabla f(u), v - u \rangle + \frac{L_s}{2}\|v - u\|_2^2. \tag{19}$$

This is a standard consequence of Lipschitz continuity of the gradient.

Choosing $v = u - \frac{1}{L_s}\nabla f(u)$ in Eq. (19) gives

$$f\left(u - \frac{1}{L_s}\nabla f(u)\right) \le f(u) - \frac{1}{2L_s}\|\nabla f(u)\|_2^2.$$

Since $\inf_v f(v) \le f(u - \frac{1}{L_s}\nabla f(u))$, we obtain

$$f(u) - \inf_v f(v) \ge \frac{1}{2L_s}\|\nabla f(u)\|_2^2. \tag{20}$$

By Lemma 5, each L_i and L_{sh} is $3/\tau^2$-smooth, and hence also $4/\tau^2$-smooth. Therefore, Eq. (20) applies to each L_i and L_{sh} with $L_s = 4/\tau^2$. Thus, Eq. (20) specializes to

$$f(u) - \inf_v f(v) \ge \frac{\tau^2}{8}\|\nabla f(u)\|_2^2, \qquad f \in \{L_1, \dots, L_N, L_{\mathrm{sh}}\}. \tag{21}$$

We now apply Eq. (21) to the two strategy-space relaxations in Proposition 1.

From a fixed rule to a shared combination strategy. Theorem 2 assumes that the optimal population loss of the fixed rule is attained at the finite vector $w_F^* = \alpha^* \mathbf{e}_F$. Assigning the same vector to all agents, the definition of L_{sh} gives

$$L_{\mathrm{sh}}(w_F^*) = \frac{1}{N}\sum_{i=1}^{N} L_i(w_F^*) = L_F^*. \tag{22}$$

This equality holds because $(w_F^*, \dots, w_F^*)$ is precisely the optimal point in the fixed-rule class specified in the theorem. Since $L_{\mathrm{sh}}^* = \inf_w L_{\mathrm{sh}}(w)$, applying Eq. (21) to $f = L_{\mathrm{sh}}$ at $u = w_F^*$ gives $L_{\mathrm{sh}}(w_F^*) - L_{\mathrm{sh}}^* \ge \frac{\tau^2}{8}\left\|\nabla L_{\mathrm{sh}}(w_F^*)\right\|_2^2$. Using Eq. (22), we obtain

$$L_F^* - L_{\mathrm{sh}}^* \ge \frac{\tau^2}{8}\left\|\nabla L_{\mathrm{sh}}(w_F^*)\right\|_2^2. \tag{23}$$

From a shared combination strategy to personalized strategies. By the definition of $L(W)$, the personalized population objective is separable with respect to the N agent-specific vectors:

$$L_{\mathrm{per}}^* = \inf_{w_1, \dots, w_N} \frac{1}{N}\sum_{i=1}^{N} L_i(w_i). \tag{24}$$

We now show explicitly that the infimum of this separable sum equals the sum of the individual infima. For any parameter tuple $(w_1, \dots, w_N)$, $L_i(w_i) \ge \inf_w L_i(w)$ for every i. Therefore,

$$L_{\mathrm{per}}^* \ge \frac{1}{N}\sum_{i=1}^{N} \inf_w L_i(w). \tag{25}$$

Conversely, let $\xi > 0$ be arbitrary. By the definition of the infimum, for every agent i, there exists a vector $w_i^{(\xi)}$ such that $L_i\left(w_i^{(\xi)}\right) \le \inf_w L_i(w) + \xi$. Substituting the particular tuple

$\left(w_1^{(\xi)}, \ldots, w_N^{(\xi)}\right)$ into Eq. (24) gives $L^*_{\mathrm{per}} \le \frac{1}{N}\sum_{i=1}^N \inf_w L_i(w) + \xi$. Since this inequality holds for every $\xi > 0$, letting $\xi \downarrow 0$ and combining it with Eq. (25) yields the exact separability identity

$$L^*_{\mathrm{per}} = \frac{1}{N}\sum_{i=1}^N \inf_w L_i(w). \tag{26}$$

This argument does not require the personalized infimum of each L_i to be attained by any finite parameter vector.

By assumption, the optimum of the shared strategy is attained at the finite vector w^*_{sh}. Therefore,

$$L^*_{\mathrm{sh}} = L_{\mathrm{sh}}(w^*_{\mathrm{sh}}) = \frac{1}{N}\sum_{i=1}^N L_i(w^*_{\mathrm{sh}}). \tag{27}$$

For each agent i, applying Eq. (21) to $f = L_i$ at the common point $u = w^*_{\mathrm{sh}}$ gives

$$L_i(w^*_{\mathrm{sh}}) - \inf_w L_i(w) \ge \frac{\tau^2}{8}\left\|\nabla L_i(w^*_{\mathrm{sh}})\right\|_2^2. \tag{28}$$

Averaging Eq. (28) over i and substituting Eqs. (26) and (27), we obtain

$$L^*_{\mathrm{sh}} - L^*_{\mathrm{per}} \ge \frac{\tau^2}{8N}\sum_{i=1}^N \left\|\nabla L_i(w^*_{\mathrm{sh}})\right\|_2^2. \tag{29}$$

Combining the two gains. The overall difference admits the following exact telescoping decomposition: $L^*_F - L^*_{\mathrm{per}} = (L^*_F - L^*_{\mathrm{sh}}) + \left(L^*_{\mathrm{sh}} - L^*_{\mathrm{per}}\right)$. Substituting Eq. (23) and Eq. (29) into this identity yields

$$L^*_F - L^*_{\mathrm{per}} \ge \frac{\tau^2}{8}\left\|\nabla L_{\mathrm{sh}}(w^*_F)\right\|_2^2 + \frac{\tau^2}{8N}\sum_{i=1}^N \left\|\nabla L_i(w^*_{\mathrm{sh}})\right\|_2^2, \tag{30}$$

which is exactly the result stated in Theorem 2. The interpretation given in the main text also follows directly from the nonnegativity of squared norms. If $\nabla L_{\mathrm{sh}}(w^*_F) \neq 0$, then the first term in Eq. (30) is strictly positive, implying a strict reduction in the population loss when relaxing the fixed rule to a shared combination strategy. If there exists at least one agent i such that $\nabla L_i(w^*_{\mathrm{sh}}) \neq 0$, then the second term is strictly positive, showing that a single shared vector cannot simultaneously attain the individual optimum for all agents, and personalization further strictly reduces the population loss.

□

D.3 Proof of theorem 3

Proof. We prove the theorem by bounding the finite-step empirical optimization error and the finite-sample generalization error, and then combining the two bounds.

Finite-Step Empirical Optimization Error of Full-Batch Gradient Descent. For agent i, define $\widehat{L}_i(w) = \frac{1}{|\mathcal{D}_{\mathrm{sel}}|}\sum_{(x,y^*)\in\mathcal{D}_{\mathrm{sel}}} \ell_i^{\mathrm{sel}}(x, y^*; w)$. Accordingly, $\mathcal{L}_{\mathrm{sel}}(W) = \frac{1}{N}\sum_{i=1}^N \widehat{L}_i(w_i)$, so the empirical objective is separable across agents.

Fix an agent i. Starting from w_i^0, full-batch gradient descent performs

$$w_i^{s+1} = w_i^s - \eta\nabla\widehat{L}_i(w_i^s), \qquad s = 0, \ldots, S-1, \tag{31}$$

where $\eta \le \tau^2/3$. Let $\widehat{w}_i = w_i^S$ and $q = \eta S$. By Lemma 5, $\widehat{L}_i$ is convex and L_s-smooth with $L_s = \frac{3}{\tau^2}$, and hence $\eta \le 1/L_s$.

For any comparator u, convexity and the Descent Lemma give

$$\widehat{L}_i(w_i^{s+1}) - \widehat{L}_i(u) \le \left\langle \nabla\widehat{L}_i(w_i^s), w_i^s - u\right\rangle - \frac{\eta}{2}\left\|\nabla\widehat{L}_i(w_i^s)\right\|_2^2.$$

Using Eq. (31), the right-hand side equals

$$\frac{\|w_i^s - u\|_2^2 - \|w_i^{s+1} - u\|_2^2}{2\eta},$$

and therefore

$$\widehat{L}_i(w_i^{s+1}) - \widehat{L}_i(u) \le \frac{\|w_i^s - u\|_2^2 - \|w_i^{s+1} - u\|_2^2}{2\eta}. \tag{32}$$

Since $\eta \le 1/L_s$, the Descent Lemma also gives $\widehat{L}_i(w_i^{s+1}) \le \widehat{L}_i(w_i^s)$. Summing Eq. (32) over $s = 0, \dots, S-1$ and using this monotonicity yields

$$\widehat{L}_i(w_i^S) - \widehat{L}_i(u) \le \frac{\|w_i^0 - u\|_2^2}{2\eta S}.$$

Taking $u = w_{i,\mathrm{per}}^*$ and using $\widehat{w}_i = w_i^S$ and $q = \eta S$, we obtain

$$\widehat{L}_i(\widehat{w}_i) - \widehat{L}_i(w_{i,\mathrm{per}}^*) \le \frac{\|w_{i,\mathrm{per}}^* - w_i^0\|_2^2}{2q}. \tag{33}$$

Note that $w_{i,\mathrm{per}}^*$ is used only as a fixed comparator and need not minimize the empirical loss $\widehat{L}_i$. Averaging over all agents gives

$$\frac{1}{N}\sum_{i=1}^{N}\left[\widehat{L}_i(\widehat{w}_i) - \widehat{L}_i(w_{i,\mathrm{per}}^*)\right] \le \frac{B_{\mathrm{per}}^2}{2q}. \tag{34}$$

Finite-Sample Generalization Error. We next establish uniform stability of the S-step full-batch gradient descent procedure. Fix an agent i, and consider two training sets $\mathcal{D}_{\mathrm{sel}}$ and $\mathcal{D}'_{\mathrm{sel}}$ of the same size that differ in exactly one valid problem. Let $\widehat{L}_i$ and $\widehat{L}'_i$ denote their empirical losses, and let w_i^s and $\widetilde{w}_i^s$ be the corresponding gradient descent iterates initialized at the same w_i^0.

For any convex L_s-smooth function f, the Baillon–Haddad inequality implies that the gradient map $G_f(w) = w - \eta\nabla f(w)$ is non-expansive whenever $\eta \le 2/L_s$. Here, $L_s = 3/\tau^2$ and $\eta \le \tau^2/3 = 1/L_s < 2/L_s$, so

$$\|G_f(u) - G_f(v)\|_2 \le \|u - v\|_2. \tag{35}$$

Since the two datasets differ in one sample and Lemma 5 gives $\|\nabla \ell_i^{\mathrm{sel}}(x, y^*; w)\|_2 \le 4/\tau$, for any w,

$$\left\|\nabla\widehat{L}_i(w) - \nabla\widehat{L}'_i(w)\right\|_2 \le \frac{8}{\tau|\mathcal{D}_{\mathrm{sel}}|}. \tag{36}$$

Applying Eqs. (35) and (36) to the two gradient updates gives

$$\|w_i^{s+1} - \widetilde{w}_i^{s+1}\|_2 \le \|w_i^s - \widetilde{w}_i^s\|_2 + \frac{8\eta}{\tau|\mathcal{D}_{\mathrm{sel}}|}.$$

Because $w_i^0 = \widetilde{w}_i^0$, iterating this inequality for S steps yields

$$\|w_i^S - \widetilde{w}_i^S\|_2 \le \frac{8q}{\tau|\mathcal{D}_{\mathrm{sel}}|}. \tag{37}$$

Again by Lemma 5, ℓ_i^{sel} is $4/\tau$-Lipschitz. Hence, for any valid test problem $(x_{\mathrm{test}}, y_{\mathrm{test}}^*)$,

$$\left|\ell_i^{\mathrm{sel}}(x_{\mathrm{test}}, y_{\mathrm{test}}^*; w_i^S) - \ell_i^{\mathrm{sel}}(x_{\mathrm{test}}, y_{\mathrm{test}}^*; \widetilde{w}_i^S)\right| \le \frac{32q}{\tau^2|\mathcal{D}_{\mathrm{sel}}|}. \tag{38}$$

Thus, the learning algorithm is uniformly stable with

$$\beta = \frac{32q}{\tau^2|\mathcal{D}_{\mathrm{sel}}|}.$$

By the standard expected-generalization result for uniformly stable algorithms (Hardt et al., 2016),

$$\left|\mathbb{E}_{\mathcal{D}_{\mathrm{sel}}}\left[L_i(\widehat{w}_i) - \widehat{L}_i(\widehat{w}_i)\right]\right| \le \frac{32q}{\tau^2|\mathcal{D}_{\mathrm{sel}}|}. \tag{39}$$

In particular,

$$\mathbb{E}_{\mathcal{D}_{\mathrm{sel}}}\left[L_i(\widehat{w}_i) - \widehat{L}_i(\widehat{w}_i)\right] \le \frac{32q}{\tau^2|\mathcal{D}_{\mathrm{sel}}|}.$$

Combining the Optimization and Generalization Errors. For each agent,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[L_i(\widehat{w}_i)\right] - L_i(w^*_{i,\text{per}})\\ &= \mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[L_i(\widehat{w}_i) - \widehat{L}_i(\widehat{w}_i)\right] + \mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[\widehat{L}_i(\widehat{w}_i) - \widehat{L}_i(w^*_{i,\text{per}})\right]\\ &\quad + \mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[\widehat{L}_i(w^*_{i,\text{per}}) - L_i(w^*_{i,\text{per}})\right].\end{aligned} \tag{40}$$

The first term is bounded by Eq. (39), and the second by Eq. (33). For the third term, $w^*_{i,\text{per}}$ is fixed independently of $\mathcal{D}_{\text{sel}}$, and the samples are i.i.d., so

$$\mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[\widehat{L}_i(w^*_{i,\text{per}})\right] = L_i(w^*_{i,\text{per}}).$$

Therefore,

$$\mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[L_i(\widehat{w}_i)\right] - L_i(w^*_{i,\text{per}}) \le \frac{\|w^*_{i,\text{per}} - w^0_i\|_2^2}{2q} + \frac{32q}{\tau^2|\mathcal{D}_{\text{sel}}|}.$$

Averaging over all agents and using the definition of B^2_{per} gives

$$\mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[L(\widehat{W})\right] - L(W^*_{\text{per}}) \le \frac{B^2_{\text{per}}}{2q} + \frac{32q}{\tau^2|\mathcal{D}_{\text{sel}}|}. \tag{41}$$

Since W^*_{per} attains the optimal personalized population loss, $L(W^*_{\text{per}}) = L^*_{\text{per}}$. Hence,

$$\mathbb{E}_{\mathcal{D}_{\text{sel}}}\left[L(\widehat{W})\right] - L^*_{\text{per}} \le \frac{B^2_{\text{per}}}{2q} + \frac{32q}{\tau^2|\mathcal{D}_{\text{sel}}|},$$

□

D.4 Detailed Derivation of Remark

Sample-efficient personalized strategy learning. For $B_{\text{per}} > 0$, define

$$\psi(q) = \frac{B^2_{\text{per}}}{2q} + \frac{32q}{\tau^2|\mathcal{D}_{\text{sel}}|}. \tag{42}$$

Its first and second derivatives are

$$\psi'(q) = -\frac{B^2_{\text{per}}}{2q^2} + \frac{32}{\tau^2|\mathcal{D}_{\text{sel}}|}, \qquad \psi''(q) = \frac{B^2_{\text{per}}}{q^3} > 0. \tag{43}$$

Thus, ψ is strictly convex on $q > 0$, and its unique stationary point is the global minimizer. Solving $\psi'(q) = 0$ gives

$$q^* = \frac{\tau B_{\text{per}}\sqrt{|\mathcal{D}_{\text{sel}}|}}{8}. \tag{44}$$

At $q = q^*$, the two terms in Eq. (42) are equal:

$$\frac{B^2_{\text{per}}}{2q^*} = \frac{32q^*}{\tau^2|\mathcal{D}_{\text{sel}}|} = \frac{4B_{\text{per}}}{\tau\sqrt{|\mathcal{D}_{\text{sel}}|}}.$$

Substituting q^* into Eq. (42) therefore yields

$$\mathbb{E}\left[L(\widehat{W})\right] - L^*_{\text{per}} \le \frac{8B_{\text{per}}}{\tau\sqrt{|\mathcal{D}_{\text{sel}}|}} = O\Big(|\mathcal{D}_{\text{sel}}|^{-1/2}\Big). \tag{45}$$

The choice in Eq. (44) is compatible with the step-size condition in Theorem 3. Since $q = \eta S$, for any prescribed $q^* > 0$, it suffices to choose a finite integer $S \ge \left\lceil \frac{3q^*}{\tau^2} \right\rceil$ and set $\eta = q^*/S$. Then $\eta \le \tau^2/3$ and $\eta S = q^*$. Hence, the minimizing value q^* can be realized by a finite number of gradient-descent steps.

If $B_{\text{per}} = 0$, then by definition $w^*_{i,\text{per}} = w^0_i$ for every agent i. Thus, the initialization already coincides with the optimal personalized strategy, and no further adaptation is required.

Finally, to achieve an expected excess population strategy loss of at most $\epsilon > 0$, Eq. (45) shows that it is sufficient to require $\frac{8B_{\mathrm{per}}}{\tau\sqrt{|\mathcal{D}_{\mathrm{sel}}|}} \le \epsilon$, or equivalently,

$$|\mathcal{D}_{\mathrm{sel}}| \ge \frac{64B_{\mathrm{per}}^2}{\tau^2\epsilon^2}. \tag{46}$$

Therefore, for fixed B_{per} and τ, the sufficient sample complexity for achieving ϵ-near-optimal population strategy loss is $O(\epsilon^{-2})$. Since Eq. (6) optimizes only the four-dimensional strategy parameter w_i of each agent while the underlying LLMs remain frozen, this sample complexity has no explicit dependence on the parameter size of the underlying LLMs. This provides the theoretical basis: SocialMAD can approach the optimal information-utilization behavior within the strategy space from limited feedback by learning only personalized combinations of four social learning units.

Explicit personalization gains under finite samples. Let $\Delta := L_{\mathrm{sh}}^* - L_{\mathrm{per}}^* > 0$. Combining this population-level gain with the finite-sample error in Eq. (45) gives

$$L_{\mathrm{sh}}^* - \mathbb{E}\Big[L(\widehat{W})\Big] \ge \Delta - \frac{8B_{\mathrm{per}}}{\tau\sqrt{|\mathcal{D}_{\mathrm{sel}}|}}. \tag{47}$$

Thus, finite-sample learning preserves the personalization benefit whenever the estimation error is smaller than the population gain. In particular, if

$$|\mathcal{D}_{\mathrm{sel}}| > \frac{64B_{\mathrm{per}}^2}{\tau^2\Delta^2}, \tag{48}$$

the learned personalized strategy strictly outperforms the shared optimum in expectation.

Moreover, if the finite-sample error is at most $\Delta/2$, then

$$L_{\mathrm{sh}}^* - \mathbb{E}\Big[L(\widehat{W})\Big] \ge \frac{\Delta}{2}, \tag{49}$$

so at least half of the personalization gain is retained in expectation. This is guaranteed when

$$|\mathcal{D}_{\mathrm{sel}}| \ge \frac{256B_{\mathrm{per}}^2}{\tau^2\Delta^2}. \tag{50}$$

Since Proposition 1 gives $L_{\mathrm{sh}}^* \le L_F^*$, the same expected advantage also holds over the optimal fixed strategy.

Finally, Theorem 2 gives

$$L_{\mathrm{sh}}^* - L_{\mathrm{per}}^* \ge \frac{\tau^2}{8N}\sum_{i=1}^{N}\|\nabla L_i(w_{\mathrm{sh}}^*)\|_2^2. \tag{51}$$

Hence, whenever at least one agent has $\nabla L_i(w_{\mathrm{sh}}^*) \neq 0$, personalization provides a strict population-level gain, which can be retained in expectation with sufficient finite-sample feedback.

Theoretical basis for judgment preservation and error correction. Fix an agent i and an arbitrary strategy parameter w_i. The personalized strategy selects the current social reference by $k_i = \arg\max_{1\le k\le K}\pi_i(k \mid p_i, \mathcal{G})$, where k^* denotes the index of the correct candidate.

Suppose that $k_i \neq k^*$. Since k_i maximizes the strategy preference, $\pi_i(k_i \mid p_i, \mathcal{G}) \ge \pi_i(k^* \mid p_i, \mathcal{G})$. Because k_i and k^* are distinct and the candidate probabilities sum to one, it follows that

$$\pi_i(k^* \mid p_i, \mathcal{G}) \le \frac{1}{2}. \tag{52}$$

The corresponding single-sample strategy loss is $\ell_i^{\mathrm{sel}}(x, y^*; w_i) = -\log\pi_i(k^* \mid p_i, \mathcal{G})$. Therefore, whenever $k_i \neq k^*$, Eq. (52) implies $\ell_i^{\mathrm{sel}}(x, y^*; w_i) \ge \log 2$. Since the loss is nonnegative when $k_i = k^*$, we have the pointwise bound

$$\mathbb{I}\{k_i \neq k^*\} \le \frac{\ell_i^{\mathrm{sel}}(x, y^*; w_i)}{\log 2}. \tag{53}$$

Taking expectation over valid initial problems gives

$$\Pr(k_i \neq k^*) \leq \frac{L_i(w_i)}{\log 2}. \tag{54}$$

Thus, the population strategy loss directly controls the probability of selecting an incorrect social reference. For the learned parameter $\widehat{w}_i$, Eq. (54) holds conditionally on every realization of the training set. Taking expectation over $\mathcal{D}_{\mathrm{sel}}$ and averaging over agents yields

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}_{\mathrm{sel}}}\left[\Pr(k_i \neq k^* \mid \widehat{w}_i)\right] \leq \frac{\mathbb{E}_{\mathcal{D}_{\mathrm{sel}}}\left[L(\widehat{W})\right]}{\log 2}. \tag{55}$$

Combining Eq. (55) with Theorem 3 further gives

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}_{\mathrm{sel}}}\left[\Pr(k_i \neq k^* \mid \widehat{w}_i)\right] \leq \frac{1}{\log 2}\left(L^*_{\mathrm{per}} + \frac{B^2_{\mathrm{per}}}{2q} + \frac{32q}{\tau^2|\mathcal{D}_{\mathrm{sel}}|}\right). \tag{56}$$

Hence, finite-sample strategy learning also provides an explicit bound on the expected correct-reference selection error.

We next distinguish the two agent states central to collaborative reasoning. By the law of total probability,

$$\begin{aligned}\Pr(k_i \neq k^*) = {} & \Pr(y_i^0 = y^*)\Pr(k_i \neq k^* \mid y_i^0 = y^*) \\ & + \Pr(y_i^0 \neq y^*)\Pr(k_i \neq k^* \mid y_i^0 \neq y^*).\end{aligned} \tag{57}$$

The first term measures reference-selection errors for agents whose initial judgments are already correct and is therefore associated with correct-judgment preservation. The second measures reference-selection errors for initially incorrect agents and is associated with error correction. Thus, Eqs. (54) and (57) show that the strategy loss jointly controls the state-weighted reference-selection risks underlying these two behaviors. Since reference information quality is closely related to subsequent revision and final performance (Tian et al., 2026), this reference-selection guarantee provides an indirect theoretical basis for improved judgment preservation and error correction.

E Additional Experimental Details

E.1 Benchmarks

We evaluate SocialMAD on eight reasoning benchmarks spanning arithmetic word problems, competition-level mathematics, and logical reasoning.

For arithmetic word-problem solving, we use the GSM8K test set (Cobbe et al., 2021), containing 1,319 problems, and SVAMP (Patel et al., 2021), containing 1,000 problems with varied linguistic and structural forms.

For competition-level mathematical reasoning, we use MATH-500, a 500-problem evaluation subset of MATH (Hendrycks et al., 2021b; Lightman et al., 2024), together with AIME 2024 and AIME 2025, each containing 30 problems from the corresponding AIME competitions.

For logical reasoning, we use FOLIO-Wiki (Han et al., 2024; Wang et al., 2024), a curated 460-example subset drawn from the WikiLogic portion of FOLIO with three-way labels of *True*, *False*, and *Unknown*; MMLU-Logic (Hendrycks et al., 2021a), the 126-question test split of the formal logic subject in MMLU; and PrOntoQA (Saparov & He, 2023), for which we adopt the 500-example 5-hop subset of the fictional-characters setting for binary deductive reasoning.

E.2 LLMs and Inference Settings

LLM Configurations. All experiments use the officially released instruction-tuned checkpoints of the evaluated LLMs. In addition, we further evaluate a 7B same-family setting using the corresponding Qwen2.5-7B-Instruct, Qwen2.5-Math-7B-Instruct, and Qwen2.5-Coder-7B-Instruct models (Qwen et al., 2025; Hui et al., 2024; Yang et al., 2024). All backbone LLMs remain frozen during

both strategy optimization and inference. The model checkpoints are initialized with the released pretrained weights and no additional fine-tuning is performed. For generation, we use stochastic decoding with temperature 1.0 for most models. For Qwen2.5-Math-7B-Instruct, the temperature is set to 0.2 to avoid unstable generation under higher sampling temperatures. Unless otherwise specified, all experiments use a context length of 4096 tokens and two debate rounds.

Evaluation Protocol. For fair comparison, all methods use identical initial agent responses under the same random seed. During evaluation, generated answers are matched using semantic equivalence rather than exact string matching. This evaluation protocol is applied uniformly to all methods, ensuring that the improvements are not caused by different answer matching criteria. For LASE, we evaluate each agent as the leader and average over all leader choices.

Computational Resources. All experiments are conducted on RTX A6000 GPUs.

Confidence Estimation. For each generated response $m_i = (y_i, r_i, \rho_i)$, we estimate the confidence score ρ_i from the model's token distributions. Let $(\omega_{i,1}, \ldots, \omega_{i,H_i})$ denote the H_i tokens of r_i. We compute $\rho_i = \exp\left(\frac{1}{H_i}\sum_{h=1}^{H_i} \log \max_{\omega\in\mathcal{V}} \Pr_{G_i}(\omega \mid \omega_{i,<h}, x)\right), \rho_i \in (0, 1]$, where $\Pr_{G_i}$ denotes the token probability assigned by agent i's generation model. Higher scores indicate greater generation confidence.

E.3 Strategy Optimization

We optimize only the personalized social strategy parameters while keeping all LLM backbones frozen. For all agent settings, the strategy optimization follows the same offline training procedure. For the 1.5B same-family setting, we randomly sample 100 training examples from the GSM8K training set for strategy optimization. For the 7B same-family and cross-family settings, we randomly sample 100 examples from the MATH training set. We then construct D_{sel} from these examples based on the agents' responses, retaining only examples with at least one correct candidate. The strategy parameters are optimized for 3 epochs using a residual-form update with coefficient 0.1 and a temperature of 0.3, corresponding to a cross-entropy gradient step size of $\eta = 0.03$. After offline optimization, the learned strategies are fixed and directly evaluated on all benchmarks without any additional target-task training or tuning.

E.4 Prompt Templates

Initial Reasoning Prompt. Following the standard MAD protocol (Du et al., 2024), all agents independently generate their initial responses using the following instruction:

```
Please solve the problem step by step.
```

Revision Prompt. In subsequent interaction rounds, all methods use the same revision prompt:

```
These are the potential solutions to the problem:{context}
Use the potential solutions as additional information for the
following question.
Question:{question}
Please think step by step and solve the problem.
```

For vanilla MAD, `context` contains the raw responses from all agents. For SocialMAD, `context` contains the candidate cluster selected by the personalized social strategy.

E.5 Evaluation Details

Accuracy. For all benchmarks, we report the standard accuracy:

$$\text{Accuracy} = \frac{\sum \mathbb{I}(y^{\text{final}} \sim_x y^*)}{\sum 1},$$

where y^{final} and y^* denote the final prediction and the ground-truth answer, respectively.

Preservation and Correction. To evaluate how agents revise their initial judgments after social interaction, we measure correction and preservation based on the transition between the initial judgment y^{init} and the final judgment y^{final}.

Preservation measures the fraction of initially correct judgments that remain correct after revision:

$$\text{Preservation} = \frac{\sum \mathbb{I}(y^{\text{init}} \sim_x y^*, y^{\text{final}} \sim_x y^*)}{\sum \mathbb{I}(y^{\text{init}} \sim_x y^*)}.$$

Correction measures the fraction of initially incorrect judgments that are corrected after revision:

$$\text{Correction} = \frac{\sum \mathbb{I}(y^{\text{init}} \not\sim_x y^*, y^{\text{final}} \sim_x y^*)}{\sum \mathbb{I}(y^{\text{init}} \not\sim_x y^*)}.$$

For the majority–minority conflict analysis, we partition three-agent cases according to the correctness composition of the initial judgments. Correct-minority preservation is computed on cases with exactly one initially correct agent and measures whether this agent remains correct after interaction. Incorrect-minority correction is computed on cases with exactly one initially incorrect agent and measures whether this agent becomes correct after interaction.

Wrong Adoption Rate. To evaluate robustness against repeated erroneous rationales, we measure the wrong adoption rate under the constructed erroneous-context setting. It quantifies the fraction of agents whose final judgments are changed to the repeated erroneous answer among those whose initial judgments do not match that erroneous answer:

$$\text{Wrong Adoption} = \frac{\sum \mathbb{I}(y^{\text{final}} \sim_x y^{\text{err}}, y^{\text{init}} \not\sim_x y^{\text{err}})}{\sum \mathbb{I}(y^{\text{init}} \not\sim_x y^{\text{err}})},$$

where y^{err} denotes the erroneous answer repeatedly injected into the social context.

Cluster Selection and Revision Metrics. To analyze the learned social strategy, we evaluate whether the selected candidate cluster contains the correct answer and whether the subsequent revision effectively utilizes the selected information. We compare the index k_i of the cluster selected by agent a_i with the correct cluster index k^*, whose candidate answer matches the ground-truth answer. Because selecting the correct reference does not necessarily lead to a correct final response due to the subsequent LLM reasoning process, these metrics are used only to diagnose strategy selection, while final task accuracy measures end-to-end performance.

The correct-cluster selection accuracy is defined as:

$$\text{Selection Accuracy} = \frac{\sum \mathbb{I}(k_i = k^*)}{\sum 1}.$$

Correct-cluster preservation measures whether the agent i whose initial judgment is correct continues to select a cluster containing the correct answer:

$$\text{Correct-cluster Preservation} = \frac{\sum \mathbb{I}\left(y^{\text{init}} \sim_x y^*,\ k_i = k^*\right)}{\sum \mathbb{I}\left(y^{\text{init}} \sim_x y^*\right)}.$$

Wrong-cluster recovery measures whether the agent i whose initial judgment is incorrect switches to a cluster containing the correct answer:

$$\text{Wrong-cluster Recovery} = \frac{\sum \mathbb{I}\left(y^{\text{init}} \not\sim_x y^*,\ k_i = k^*\right)}{\sum \mathbb{I}\left(y^{\text{init}} \not\sim_x y^*\right)}.$$

E.6 Experimental Variants and Ablation Settings

Social Information Representation Variants. Fig. 4 compares four social information representations. *MAD* uses the standard response-level representation, directly providing the raw responses from the previous round as the social context for subsequent reasoning. *Simple Dedup* randomly retains one supporting rationale among duplicated responses and removes the others, reducing repeated presentation of identical or superficially redundant information. *Structured w/o Evidence Consolidation* groups semantically equivalent responses into candidate clusters, shifting the interaction unit from individual responses to distinct candidate viewpoints, but does not consolidate

evidence within each cluster. *Structured (Ours)* further applies Evidence Consolidation to organize the reasoning evidence within each candidate cluster, preserving useful supporting evidence. These progressively constructed variants isolate the contribution of the complete structured social information representation and its key components to reducing information interference and supporting effective reasoning.

In Fig. 4(c), we further evaluate the robustness of these representations to repeated errors on the subset of MATH-500 examples containing at least one erroneous rationale. To isolate representation, personalized strategy selection is disabled for variants, and the resulting representation is directly supplied to the common revision context. Specifically, the same erroneous rationale is additionally repeated $1, 2, 4, 8$ times in the social context, increasing the repetition of the same erroneous viewpoint and testing whether different representations treat repeated instances of the same error as multiple independent pieces of evidence.

Social Learning Strategy Variants. Fig. 5(a,b) compares four fixed strategies, a shared strategy, and the personalized strategy. Each fixed strategy applies a single social learning behavior uniformly to all agents. *Self-persistence* selects a candidate cluster based only on the consistency with the agent's current judgment, $\iota_{i,k}$, and therefore favors the candidate cluster consistent with its own judgment. *Copy-when-uncertain* uses only $(1 - \rho_i)(1 - \iota_{i,k})$, assigning higher preference to alternative candidates as the agent's confidence decreases. *Conformity* selects candidates based only on relative group support Δn_k, favoring viewpoints supported by more agents. *Confidence-based following* selects candidates based only on relative group confidence $\Delta\bar{\rho}_k$, favoring viewpoints with higher group confidence.

Shared Policy learns a single shared strategy parameter $w \in \mathbb{R}^4$ for all agents. It can combine multiple social learning behaviors according to the current individual state and social information, but cannot capture agent-specific differences in strategy preference. *SocialMAD* instead learns an agent-specific strategy parameter w_i for each agent, enabling personalized social learning strategies.

Strategy Feature Ablation. Fig. 5(c) removes one strategy feature from Eq. (3) at a time to examine the roles of individual-state and social cues in personalized strategy learning. Specifically, *w/o Consistency* removes the consistency feature $\iota_{i,k}$; *w/o Uncertainty* removes the uncertainty-dependent feature $(1 - \rho_i)(1 - \iota_{i,k})$; *w/o Support* removes the relative group-support feature Δn_k; and *w/o Group-confidence* removes the relative group-confidence feature $\Delta\bar{\rho}_k$. For each ablation, we retrain the strategy parameters with the remaining training settings unchanged and use the retrained strategy for inference. Fig. 5(c) reports the resulting performance changes relative to the full four-feature strategy, measuring the contribution of each social learning cue.

Scaling Setting. In the scaling experiments in Fig. 6, we increase the number of agents by instantiating additional independent copies of the three heterogeneous LLMs used in the main setting. For $N \in \{3, 6, 9, 12\}$, we instantiate $1, 2, 3, 4$ agents per LLM, respectively, while keeping the model composition ratio unchanged. Each instance generates responses independently and optimizes its personalized strategy parameter w_i separately using the same training set. For round scaling, we fix 6 agents with two instances of each LLM, and increase the number of interaction rounds under this configuration.

F Additional Results

F.1 Comparison Experiments

7B Same-Family Setting. As shown in Table 2, we further evaluate SocialMAD under the 7B same-family setting. SocialMAD outperforms MAD on seven benchmarks and matches it on AIME25. Compared with all baselines, SocialMAD achieves the best or tied-best performance on six benchmarks and ranks second on the remaining two. These consistent improvements across mathematical and logical reasoning tasks show that the gains of SocialMAD do not rely on cross-family model diversity and remain effective when all agents come from the same model family.

Homogeneous-Agent Settings. To further evaluate the effectiveness of SocialMAD under homogeneous-agent settings, we conduct additional experiments with two homogeneous agent configurations, where all agents use the same LLM but generate independent responses through stochastic decoding. As shown in Table 3, SocialMAD consistently achieves strong performance across dif-

Table 2: Accuracy (%) across eight benchmarks under the 7B same-family agent setting. CoT and CoT-SC report the average accuracy of the three agents. For LASE, we report the average accuracy across all possible leader choices.

Method	GSM8K	MATH -500	AIME24	AIME25	SVAMP	FOLIO-Wiki	MMLU-Logic	PrOntoQA
Qwen2.5-7B / Qwen2.5-Math-7B / Qwen2.5-Coder-7B								
CoT	79.00	55.53	7.78	6.67	86.83	74.64	52.91	86.07
CoT-SC	84.88	58.40	7.78	4.44	90.50	74.71	55.82	87.93
Majority Vote	89.69	61.80	6.67	6.67	92.60	83.26	57.14	93.20
MAD	93.18	65.00	3.33	10.00	93.40	86.30	60.32	96.00
RECONCILE	94.84	64.80	10.00	6.67	94.50	85.87	56.35	93.40
MAD-Conformist	93.18	63.40	6.67	10.00	93.70	84.57	58.73	93.20
MAD-M^2	95.53	71.40	**16.67**	6.67	93.00	83.26	57.94	96.20
Sparse MAD	93.93	67.60	10.00	3.33	93.70	85.65	60.32	96.40
MasRouter	94.39	68.40	10.00	10.00	94.20	82.61	**62.70**	96.40
LASE	87.29	60.00	10.00	5.56	89.97	74.71	55.29	85.80
SocialMAD	**95.68**	**75.80**	13.33	**10.00**	**95.30**	**88.26**	61.90	**96.80**

Table 3: Accuracy (%) across two homogeneous-agent settings. CoT and CoT-SC report the average accuracy of all agents. For LASE, we report the average accuracy across all possible leader choices.

Method	GSM8K	MATH -500	AIME24	AIME25	SVAMP	FOLIO-Wiki	MMLU-Logic	PrOntoQA
Qwen2.5-Math-1.5B-Instruct × 3								
CoT	80.59	64.60	3.33	0.00	92.30	42.83	48.41	51.40
CoT-SC	85.22	69.20	10.00	0.00	94.20	45.43	52.38	52.00
Majority Vote	84.00	71.20	**13.33**	3.33	93.60	44.78	49.21	49.00
MAD	84.46	68.80	10.00	6.67	92.40	39.13	50.00	50.00
RECONCILE	85.97	70.20	10.00	0.00	94.20	41.30	51.59	52.40
MAD-Conformist	86.35	70.20	10.00	0.00	93.70	42.17	48.41	52.60
MAD-M^2	85.67	68.80	10.00	3.33	93.90	38.91	46.83	48.80
Sparse MAD	85.06	69.60	6.67	6.67	92.70	41.09	46.03	50.40
MasRouter	83.55	70.80	10.00	6.67	92.90	46.09	49.21	52.20
LASE	82.64	68.40	13.33	6.67	91.50	44.13	53.17	49.40
SocialMAD	**87.26**	**71.40**	10.00	**6.67**	**94.20**	**46.96**	**53.17**	**53.80**
Qwen2.5-Math-7B-Instruct × 3								
CoT	95.30	74.60	10.00	6.67	92.50	**51.73**	53.17	91.60
CoT-SC	95.45	73.00	6.67	6.67	93.60	50.86	57.93	94.80
Majority Vote	95.83	75.20	10.00	6.67	93.70	49.35	55.56	95.80
MAD	95.75	72.00	6.67	3.33	92.70	47.61	55.56	90.00
RECONCILE	95.91	72.80	10.00	6.67	94.10	49.57	57.14	94.80
MAD-Conformist	95.83	73.60	10.00	6.67	93.70	49.57	57.14	95.60
MAD-M^2	95.68	75.60	6.67	6.67	93.80	48.04	59.52	95.40
Sparse MAD	95.91	74.00	**13.33**	3.33	92.90	46.30	57.14	93.40
MasRouter	96.13	72.20	10.00	6.67	94.00	48.91	58.73	95.20
LASE	95.30	68.60	3.33	3.33	92.40	46.96	59.52	89.20
SocialMAD	**96.21**	**76.20**	10.00	**6.67**	**94.60**	51.30	**61.11**	**95.80**

ferent model scales. Under the Qwen2.5-Math-1.5B-Instruct setting, SocialMAD achieves 87.26% accuracy on GSM8K, improving over MAD by 2.81%. Under the Qwen2.5-Math-7B-Instruct setting, it achieves 76.20% accuracy on MATH-500, outperforming MAD by 4.20%. These results demonstrate that SocialMAD can effectively leverage the diverse responses generated by homogeneous agents to improve collaborative reasoning performance. Nevertheless, single-agent reasoning remains competitive on certain tasks. For example, under the 7B setting, CoT achieves 51.73% on FOLIO-Wiki, outperforming SocialMAD's 51.30%. This may be because some tasks can be effectively solved without complex information integration, while SocialMAD primarily optimizes strategy learning for scenarios involving conflicting or uncertain social information, resulting in relatively limited gains in such cases. In contrast, SocialMAD achieves more substantial improvements on tasks requiring complex reasoning and judgment revision by leveraging structured social information and personalized social strategies.

F.2 Analysis of Personalized Social Strategies

Learned Personalized Strategy Parameters. To examine whether different agents learn distinct information-utilization preferences, we first visualize the learned social learning strategy parameters of each agent under the cross-family setting. As shown in Fig. 7, different agents exhibit clearly distinct strategy-weight patterns. Qwen2.5-Math-7B assigns relatively large positive weights of 0.427

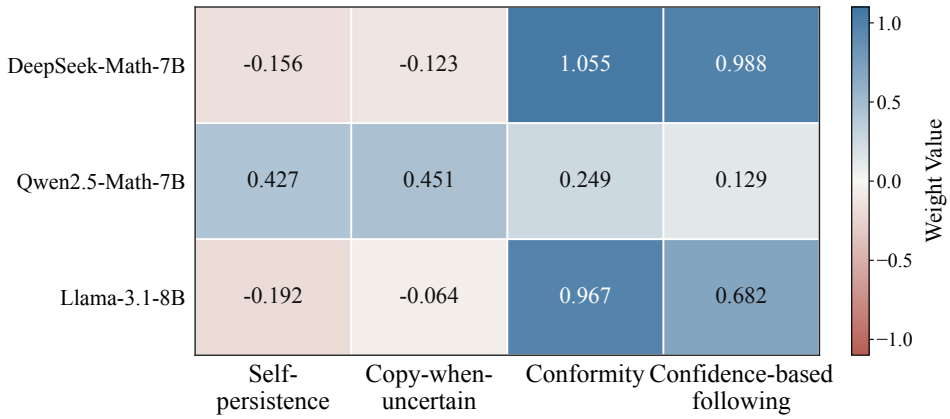


Figure 7: Learned personalized social learning strategy weights in the cross-family agent setting. Rows denote agents and columns denote the four social learning behaviors.

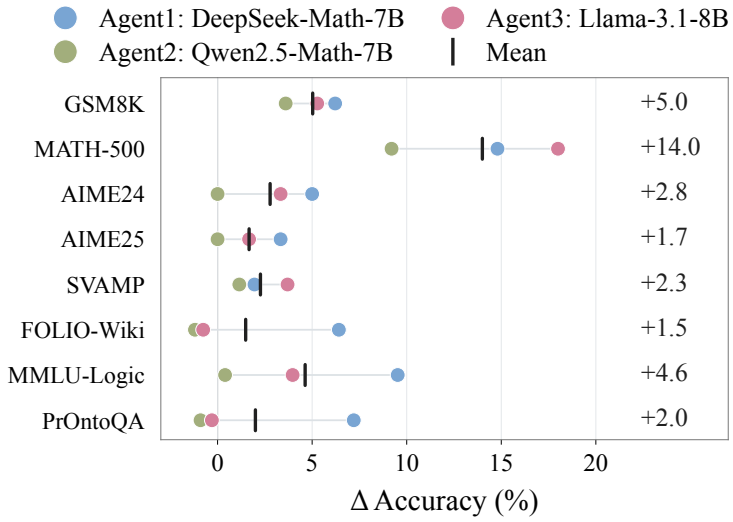


Figure 8: Cross-agent strategy swapping across eight benchmarks in the cross-family setting. For each agent, the reported value is the individual accuracy obtained with its own learned strategy minus the average accuracy obtained when using the other two agents' strategies. Colored markers denote the three agents, while black markers and the values on the right show the mean difference across agents in percentage points.

and 0.451 to self-persistence and copy-when-uncertain, respectively, while assigning lower weights to group-information-driven strategies such as conformity. This indicates that its social learning behavior is more strongly conditioned on its individual state: it tends to preserve viewpoints consistent with its current judgment, while becoming more receptive to alternative viewpoints when its own confidence is low, thereby adjusting between judgment preservation and external-information utilization.

In contrast, the weights of DeepSeek-Math-7B and Llama-3.1-8B on individual-state-related strategies are close to zero or negative. A negative weight means that activating the corresponding strategy feature decreases, rather than increases, the preference score of the candidate. For example, a negative weight on copy-when-uncertain indicates that uncertainty alone is insufficient to drive the agent toward alternative viewpoints. Instead, both agents assign substantially larger positive weights to group-information-driven strategies, indicating a stronger tendency to absorb external information

Table 4: Agent accuracy (%) under different agent strategy assignments on MATH-500. The diagonal entries correspond to each agent using its own learned strategy.

Agent / Strategy	π_1	π_2	π_3
Agent 1 DeepSeek-Math-7B	**66.00**	37.00	65.40
Agent 2 Qwen2.5-Math-7B	65.20	**74.80**	66.00
Agent 3 Llama-3.1-8B	64.80	34.00	**67.40**

based on group support and mean supporter confidence. Overall, different agents develop clearly distinct social-information utilization patterns through strategy learning, showing that SocialMAD can establish agent-specific social learning mechanisms that allow the same group information to guide reasoning revision differently for different agents.

Effect of Personalized Social Strategies. We further perform cross-agent strategy swapping to verify the necessity of learning agent-specific strategies for different agents. Fig. 8 compares the individual accuracy of each agent when using its own learned strategy versus strategies learned for other agents. Across all eight benchmarks, the average accuracy obtained with the agents' own strategies is consistently higher than that obtained with other agents' strategies, with gains of 14.0% on MATH-500, 5.0% on GSM8K, and 4.6% on MMLU-Logic. Although an individual agent may occasionally benefit from another agent's strategy on a particular task, the overall results show that replacing an agent's own strategy with one learned for another agent consistently reduces average performance, indicating that the learned strategies are specifically adapted to their corresponding agents.

Table 4 further reports the per-agent strategy-swapping results on MATH-500. All three agents achieve their highest accuracy when using their own learned strategies. In particular, Qwen2.5-Math-7B achieves 74.80% accuracy with its own strategy, which decreases to 65.20% and 66.00% when replaced by the strategies learned for the other two agents. Together with the strategy-parameter differences observed above, these results show that the strategies learned by different agents not only differ in their parameterization, but are also adapted to their respective information-utilization patterns, thereby validating the necessity of agent-specific personalization.

F.3 Analysis of Social Strategy Optimization

Sample Efficiency of Strategy Learning. We vary the number of training problems from 10 to 1000 to examine how much offline supervision is required to learn the social learning strategies. As shown in Fig. 9, correct-cluster selection accuracy improves rapidly with limited supervision and remains consistently high as the amount of training data increases. Final task accuracy shows a similar overall trend and remains comparatively stable across training sizes. These results are empirically consistent with the finite-sample analysis, showing that effective personalized social learning strategies can be learned from limited task feedback without large-scale supervision.

F.4 Token Consumption Analysis

We further analyze the token consumption of SocialMAD compared with vanilla MAD. Since both methods share the same first-round independent generation, we focus on the subsequent interaction round and report the average input-plus-output tokens consumed by all agents per sample on each benchmark. For each sample, we first sum the input and output tokens consumed by all agents during the interaction round. We then average this per-sample token consumption over all samples in each benchmark. As shown in Table 5, SocialMAD consistently reduces interaction-round token consumption while improving accuracy across all eight benchmarks. Its token cost ranges from $0.57\times$ to $0.81\times$ that of MAD, and averages only $0.70\times$ across benchmarks, corresponding to a 30.3% reduction. The reduction is particularly pronounced on MATH-500 and AIME, where SocialMAD requires only 0.57–$0.60\times$ the interaction tokens of MAD. Meanwhile, SocialMAD improves accuracy on every benchmark, with gains ranging from 7.39% to 27.00%. Overall, these results demonstrate that SocialMAD achieves more effective multi-agent reasoning with lower interaction cost.

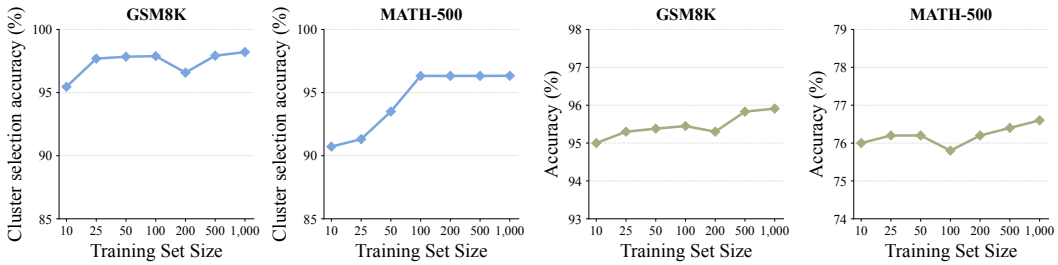


Figure 9: Effect of training-data scale on SocialMAD in the cross-family setting. The left two panels report correct-cluster selection accuracy, while the right two report final task accuracy on GSM8K and MATH-500.

Table 5: Accuracy and interaction-round token consumption of MAD and SocialMAD in the 1.5B setting. Tokens include both input and output tokens across all agents in the interaction round. Values in parentheses denote the absolute accuracy improvement over MAD and the token ratio relative to MAD, respectively.

Dataset	Accuracy (%) ↑		Tokens ↓	
	MAD	SocialMAD	MAD	SocialMAD
GSM8K	67.25	**82.26** (+15.01)	4066	**3102** (×0.76)
MATH-500	39.80	**66.80** (+27.00)	7310	**4397** (×0.60)
AIME24	0.00	**13.33** (+13.33)	9900	**5657** (×0.57)
AIME25	0.00	**10.00** (+10.00)	10369	**5859** (×0.57)
SVAMP	81.70	**93.60** (+11.90)	2726	**2207** (×0.81)
FOLIO-Wiki	53.48	**60.87** (+7.39)	4630	**3701** (×0.80)
MMLU-Logic	42.06	**55.56** (+13.50)	5621	**4124** (×0.73)
PrOntoQA	52.80	**60.40** (+7.60)	4866	**3574** (×0.73)