MMAD: Multi-Agent Mutual Awareness Debate — A Theory-of-Mind Framework for Stabilizing Small Language Model Debate

Anonymous ACL submission

Abstract

Multi-agent debate (MAD) improves reasoning in large language models but systematically *degrades* performance in small language models (SLMs, 3–8B parameters), reducing accuracy to as low as 10%, well below single-agent baselines. We identify the cause as *sycophantic drift*: agents abandon correct answers under peer pressure rather than because peers offered superior reasoning. We propose MMAD (Multi-Agent Mutual Awareness Debate), a two-tier Mutual Theory-of-Mind (MToM) framework that addresses drift at two levels: Tier 1 (peer-level MToM) requires each agent to explicitly model *why* its peers reasoned as they did before deciding whether to update; Tier 2 (teacher-level MToM) employs a 30B teacher model to perform teacher-level ToM, delivering personalised per-agent guidance that targets individual misconceptions without revealing the answer. We evaluate MMAD across four datasets spanning mathematical reasoning (GSM8K), code question answering (CodeQA, CS1QA), and commonsense reasoning (CommonsenseQA), using three SLM agents from distinct model families (Mistral-7B, Phi-4-mini, Qwen2.5-7B). MMAD achieves mean accuracy gains of +73.3 pp on GSM8K, +23.8 pp on CodeQA, +16.7 pp on CS1QA, and +31.7 pp on CommonsenseQA at Round 4 relative to single-agent baselines. Drift rates reach 0.0% by Round 4 on GSM8K, demonstrating that structured mutual awareness can effectively eliminate sycophantic capitulation in SLM debate. The source code and related materials are publicly available at `https://anonymous.4open.science/r/MMAD-0D71/`.

1 Introduction

The ability to reason correctly under uncertainty is central for deploying language models in high-stakes domains such as mathematics, programming, and scientific QA. *Multi-agent debate* (MAD) improves this capability by having multiple agents generate answers independently and iteratively refine them by inspecting each other's reasoning (Du et al., 2024).The intuition is: if independent agents agree on same conclusion, the conclusion is likely correct; if they disagree, debate helps identify the strongest reasoning.

Agent	Phase 0 (Baseline)	Round 1 (Debate)	Round 2 (Final)	Verdict
Qwen2.5 7B	(c) × Correct setup, arithmetic slip	(b) × Wrong answer	(d) Corrects itself	Improves
Mistral-7B	**(d) Correct reasoning**	**(d) Holds firm**	**(b) × Abandons answer**	**DRIFTS**
Phi-4-mini	(b) × Misreads question	(b) × Holds wrong answer	(b) × Unchanged	No change

Table 1: Reasoning drift during multi-agent debate on a math problem. *Problem: 120 new students with average age 32 join a school. The overall average age drops from 40 to 36. How many students are now in the school?* **Gold: (d) 240**

Existing work validates multi-agent debate primarily for large language models (LLMs). Du et al. (2024) show that GPT-3.5-class agents improve accuracy on arithmetic and strategic reasoning via debate. Liang et al. (2024) demonstrate that encouraging divergent thinking mitigates degeneration-of-thought, while Chen et al. (2024) introduce ReConcile, a confidence-weighted round-table framework achieving state-of-the-art results. Li et al. (2024b) further show that sparse communication topologies reduce computation without degrading debate quality. Collectively, these studies indicate that debate can enhance LLM accuracy and robustness.

Small language models (SLMs, 3–8B parameters) are increasingly important for on-device and cost-efficient deployment, and have shown strong performance on a range of tasks despite their re-

duced size (Belcak et al., 2025). However, when we apply standard multi-agent debate to SLMs on GSM8K, CodeQA, and CommonsenseQA, accuracy ranges only 10–58%, often lower than independent responses. Debate frequently degrades performance rather than improving it.

We identify this failure as **sycophantic drift**, where SLMs abandon correct answers because peers disagree. Table 1 shows agents switching to incorrect majority views after one or two rounds, even when correct reasoning was present. Prior work shows similar patterns: sequential dialogue increases capitulation (Kim and Khashabi, 2025), sycophancy limits multi-agent debate (Pitre et al., 2025), and peer confidence amplifies capitulation (Kaur, 2025). This is a robustness, not knowledge, failure: SLMs treat consensus as a proxy for correctness because they cannot reliably assess peer reasoning. The root cause is a lack of Theory of Mind (ToM)—the ability to model other agents' beliefs, intentions, and reasoning quality (Chen et al., 2025b; Wagner et al., 2025). Standard debate exposes agents only to answers or brief justifications, offering little basis to judge disagreement. Even GPT-4–class models struggle with multi-agent ToM (Li et al., 2023), though decomposing ToM into explicit sub-tasks improves higher-order reasoning (Sarangi et al., 2025).

Proposed Solution: MMAD. We propose MMAD (Multi-Agent Mutual Awareness Debate), a two-tier Mutual Theory of Mind (MToM) framework (Wang and Goel, 2022; Wang et al., 2021) designed to eliminate sycophantic drift while preserving multi-agent accuracy gains. MToM framework was proposed by Wang (Wang and Goel, 2022; Wang et al., 2021) for human–AI communication, requiring both parties to model each other's beliefs and intentions. We adapt MToM to AI–AI interactions our MMAD settings, enabling agents to model each other's knowledge, goals, and strategies to support coordinated, efficient multi-agent reasoning. MMAD follows a classroom analogy: student agents first solve problems independently, then share full reasoning traces and receive targeted teacher guidance. In group study, inspecting peers' *reasoning*—not just answers—enables principled evaluation, and when misconceptions persist, the teacher *intervenes* with questions that pinpoint the source of each agent's error.

We propose MMAD as two alternating tiers. **Tier 1: Peer-level MToM (odd rounds)** has agents inspect peers' reasoning traces and model why each peer reached its conclusion, allowing them to reject majority views based on flawed reasoning. **Tier 2: Teacher-level MToM (even rounds)** uses a 30B teacher model to perform teacher-level ToM, diagnosing each agent's errors and providing personalized guidance without revealing answers, similar to MAgICoRe (Chen et al., 2025a), but applied to metacognitive support.

We compare MMAD with a confidence-augmented debate baseline where agents report calibrated confidence scores and weight higher-confidence peers during revision (Chen et al., 2024). This baseline enables evaluation of the contribution of structured mutual awareness. We evaluate MMAD on four datasets spanning mathematical (GSM8K (Cobbe et al., 2021)), code (CodeQA (Liu and Wan, 2021), CS1QA (Lee et al., 2022)), and commonsense reasoning (CommonsenseQA (Talmor et al., 2019)) using three SLM agents from distinct model families.

Our Contributions.

1. We identify and systematically analyze *sycophantic drift* in SLM debate, where agents abandon correct answers due to peer pressure across models and tasks.
2. We introduce **MMAD**, a two-tier Mutual Theory of Mind debate framework that mitigates drift via reasoning-trace inspection and personalized teacher guidance without retraining.
3. We conduct a controlled comparison with a confidence-weighted debate baseline to isolate the effect of structured mutual awareness on answer stability.
4. Across four datasets, MMAD consistently improves accuracy and reduces drift, with the combined two-tier framework outperforming either tier individually.

2 Related Work

We frame prior work motivating MMAD in multi-agent debate, single-agent reasoning, sycophancy, and Theory of Mind, with small model reasoning informing our experiments.

Multi-Agent Debate. Researchers showed that iterative debate among LLMs improves factual and arithmetic accuracy (Du et al., 2024), and encouraging diverse reasoning mitigates degeneration-of-thought (Liang et al., 2024). Subsequent work examined how debate structure affects outcomes, such as cross-family round-table voting leverages

complementary evidence (Chen et al., 2024), sparse graphs reduce redundancy (Li et al., 2024b), and voting versus consensus benefits vary with task difficulty (Kaesberg et al., 2025). Step-level refinement with a specialist reviewer (MAgICoRe; (Chen et al., 2025a)) enhances reasoning. However MAD methods have been evaluated primarily on large models ($\geq$ 13B parameters, e.g., GPT-3.5 or GPT-4), leaving smaller models largely unexplored.

Single-Agent Reasoning and Self-Correction. MAD builds on single-agent reasoning methods such as chain-of-thought (CoT) and self-correction. CoT prompting improves performance on multi-step tasks (Wei et al., 2022), and self-consistency further boosts accuracy by aggregating multiple reasoning traces (Wang et al., 2023). However, self-correction has limitations: reported gains often disappear under fairer baselines (Kamoi et al., 2024), and iterative self-feedback can amplify model biases (Pan et al., 2024). These findings suggest that effective feedback must come from an external source, which multi-agent debate provides. In multi-agent settings, MAgICoRe (Chen et al., 2025a) shows that a stronger specialist model giving targeted feedback improves reasoning more efficiently than self-refinement or debate alone. Our Tier 2 teacher plays a similar role in MAD setting without revealing the gold answer.

Sycophancy and Answer Instability. Researchers identify sycophancy as a key reliability concern in language models: misleading prompts can flip correct answers (Rrv et al., 2024), sequential challenges increase reversals (Kim and Khashabi, 2025), and confident but incorrect peers exert disproportionate influence (Kaur, 2025). Recent work addresses sycophancy by selectively triggering debate (iMAD; (Fan et al., 2026)) or blocking unsupported updates (ConsensAgent; (Pitre et al., 2025)). MMAD complements these approaches by providing reasoning-trace evidence that helps agents evaluate peer updates independently.

Theory of Mind in Language Models. Theory of Mind (ToM) is widely studied as a benchmark of LLM cognitive depth. Li et al. (2023) show that even large models struggle to track collaborators' evolving beliefs over multi-turn interactions. Huang et al. (2024) formalise a complexity hierarchy, demonstrating that second-order ToM is harder than first-order and requires explicit simulation. Zhou et al. (2024) introduce SOTOPIA, a multi-agent environment for training socially intelligent agents with ToM reasoning.

Recent work shows that decomposing ToM into sub-steps boosts higher-order performance (Sarangi et al., 2025), and ToM capabilities scale with model size but remain unreliable below 13B parameters (Chen et al., 2025b). Our proposed MMAD operationalises a pragmatic, task-scoped form of ToM: agents are not asked to model general mental states but specifically to identify *what reasoning error* a peer made and *why that error does not justify their answer*. At Tier 2, the teacher performs teacher-level ToM over each agent's error to provide personalised guidance.

Small Language Model Reasoning. citetlu2025demystifying show that SLMs (1–8B) match LLMs on simple factual recall but lag on multi-step reasoning. Recent work attempts to close this gap via knowledge transfer, including counterfactual distillation (Feng et al., 2024), Decompose-and-Response Distillation (Li et al., 2024a), and Solution Guidance Fine-Tuning (SGFT) (Bi et al., 2025).

Unlike these methods, which require fine-tuning with gold reasoning traces, MMAD is a *training-free* inference-time framework applicable to any instruction-tuned SLM.

3 Sycophantic Drift in SLM Debate

We formally define sycophantic drift and provide empirical evidence of its prevalence across our agent ensemble.

Definition. A **drift event** for agent a at round r occurs when the agent held a correct answer at round $r-1$ and an incorrect answer at round r:

$$\text{Drift}(a, r) = \mathbf{1}\left[\text{correct}(a, r-1) \wedge \neg\,\text{correct}(a, r)\right] \tag{1}$$

The **drift rate** at round r aggregates over all agents and examples:

$$\text{DriftRate}(r) = \frac{1}{|A||D|} \sum_{a \in A} \sum_{d \in D} \text{Drift}(a, r) \tag{2}$$

where $A = \{M, P, Q\}$ and D is the evaluation dataset. We also track the **recovery rate** — the proportion of incorrect answers that become correct at the next round:

$$\text{Recovery}(a, r) = \mathbf{1}\left[\neg\,\text{correct}(a, r-1) \wedge \text{correct}(a, r)\right] \tag{3}$$

Empirical Evidence. Our preliminary experiments with standard debate show that, without

MMAD, SLMs often abandon correct answers in the first round, consistent with prior observations Pitre et al. (2025). Our MMAD framework actively minimizes DriftRate and increases RecoveryRate, producing a steadily declining drift curve alongside a rising accuracy curve.

4 The MMAD Framework

We describe MMAD's full pipeline, including the MToM tiers, prompt design, alternating round schedule, and the confidence-augmented baseline (Figure 1).

4.1 System Overview

MMAD operates in two phases:

Phase 0 — Independent Baseline. Each SLM agent independently processes the input and outputs an answer with a step-by-step reasoning trace, without inter-agent communication.

Phase 0 outputs serve both as the initial state for the first debate round and as the single-agent accuracy baseline against which all multi-agent configurations are measured.

Rounds 1–4: Alternating Debate Rounds. Agents engage in R 1 to 4 debate rounds governed by a strict alternating schedule:

- **Odd rounds** (Rounds 1, 3): *Tier 1: Peer-level MToM*. Each agent reads the full reasoning traces of both peers and must explicitly model why each peer reached their answer before deciding whether to update.
- **Even rounds** (Rounds 2, 4): *Tier 2: Teacher-level MToM*. A 30B teacher model (Qwen3-Coder-30B-A3B-Instruct) performs teacher-level ToM over all three agents' reasoning and delivers personalised, per-agent guidance targeting individual misconceptions without revealing the gold answer. Each agent then incorporates this guidance alongside peer answers to produce a revised response.

At each round, agents' produce outputs, including answers, reasoning traces, and (for the confidence baseline, confidence scores) for the next round. An independent judge model (Qwen2.5-7B-Instruct) evaluates each agent's answer at every phase and round against the gold answer, tracking per-round accuracy and drift.

4.2 Agent Models and Roles

We use three SLMs from distinct LM families as **Student Agents**: **Agent M** (`Mistral-7B-Instruct-v0.3` (Jiang et al., 2023)), **Agent P** (`Phi-4-mini-instruct` (Abdin et al., 2024)), and **Agent Q** (`Qwen2.5-7B-Instruct` (Qwen Team, 2025))—all participating in every phase and round. The **Teacher Agent**, (`Qwen3-Coder-30B-A3B-Instruct`), intervenes only in even-numbered rounds (Tier 2), privately accessing the gold answer to diagnose reasoning errors without revealing it. An independent **Judge Model**, (`Qwen2.5-7B-Instruct`), evaluates answers via semantic matching, counting paraphrase-equivalent responses as correct. Table 2 shows agents' alternating round schedule.

Round	Type	Peer	Teacher
0	Baseline (independent)	×	×
1	Tier 1 — Peer MToM		×
2	Tier 2 — Teacher MToM		
3	Tier 1 — Peer MToM		×
4	Tier 2 — Teacher MToM		

Table 2: Information available to student agents at each MMAD phase/round.

4.3 Prompt Design

Phase 0: Independent Baseline The Phase 0 prompt presents the task input and requests a dictionary: `{"predicted_output": , "explanation": }`. No inter-agent context is provided at this phase.

Tier 1: Peer-Level MToM (Odd Rounds) Each agent receives the task input, its previous answer and reasoning trace, and both peers' prior answers and traces. The key instruction is:

> *Consider the other agents' reasoning critically — they may be right or wrong. Do NOT blindly follow the majority. A lone agent can be right while two are wrong. If another agent's reasoning exposes a flaw in your answer, revise it. If you remain confident in your answer, keep it and clearly explain why.*

Tier 2: Teacher-Level MToM (Even Rounds) The teacher receives all agents' answers and reasoning traces, along with the gold answer, and outputs a per-agent guidance dictionary: `{"guide_M": , "guide_P": , "guide_Q": }`. This differs from simple answer revelation in three ways:

1. The gold answer is never copied verbatim;
2. Highlights errors to specific reasoning steps;
3. Correct answers are reinforced, e.g., "Stay confident in your answer despite peer disagreement."

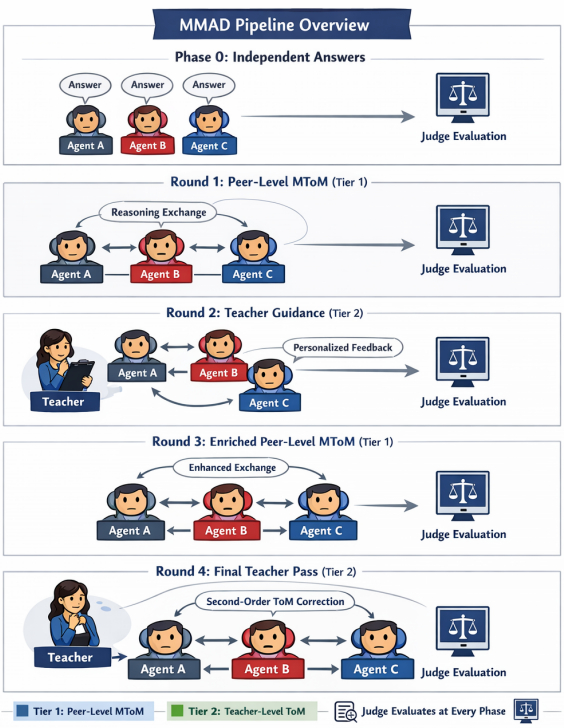

Figure 1: Overview of the MMAD pipeline. Phase 0: Three SLM agents answer independently. Round 1 (Tier 1): Agents exchange full reasoning traces via peer-level MToM. Round 2 (Tier 2): A teacher model reads all agent states and provides personalized guidance. Round 3 (Tier 1): Agents perform another peer exchange using teacher-enriched states. Round 4 (Tier 2): The teacher performs a final pass and applies teacher-level ToM correction. A judge model evaluates answer accuracy independently at every phase.

5 Experimental Setup

Implementation Details All models use `float16` precision with `device_map=auto`. Student agents generate up to 384 tokens, the teacher 512, and the judge 128, all with greedy decoding (`temperature=0`, `do_sample=False`). Models are loaded and unloaded sequentially per agent per round to prevent GPU memory issues.

Datasets. We evaluate on four benchmarks: **GSM8K** (Cobbe et al., 2021) (grade-school arithmetic word problems), **CodeQA** (Liu and Wan, 2021) (code reading comprehension), **CS1QA** (Lee et al., 2022) (introductory programming questions with open-ended answers), and **CommonsenseQA** (Talmor et al., 2019) (multiple-choice commonsense reasoning). All evaluations use representative samples from each benchmark.

Metrics. We report per-agent and mean accuracy at each round, the correct-to-incorrect flip rate (sycophantic drift rate), the incorrect-to-correct flip rate (recovery rate), and answer stability (proportion of answers unchanged from the prior round).

Baselines. All results use the single-agent baseline (Phase 0) as the primary comparison. The confidence-augmented peer-debate baseline (Conf-Debate) shares MMAD's agents, four rounds, judge, and JSON output format, but differs in two ways: (1) No teacher model is employed: all four rounds are pure peer debate; (2) Each agent reports a self-assessed confidence $c \in [0, 1]$ with its answer and weights higher-confidence peers more during revision.This design draws on ReConcile's confidence-weighted voting (Chen et al., 2024).

6 Results

Table 3 summarises per-agent and mean accuracy at Phase 0 (baseline) and Rounds 1–5 of MMAD, along with drift and recovery rates.

6.1 GSM8K

GSM8K produced the most dramatic accuracy gains, consistent with the high task difficulty for SLMs. The mean single-agent baseline accuracy was just 10.0%: Agent P (Phi-4-mini) failed entirely at baseline (0%), Agent M (Mistral-7B) achieved 5%, and Agent Q (Qwen2.5-7B) reached 25%.

MMAD improved accuracy monotonically across all rounds for every agent. After Round 4 (the second Tier 2 teacher pass), mean accuracy reached **83.3%**, a gain of +73.3 pp over baseline. Agent P showed the largest individual gain, rising from 0% to 90%; Agent M improved from 5% to 75%; Agent Q improved from 25% to 85%.

The **sycophantic drift was completely eliminated by Round 4**: all three agents recorded a drift rate of 0.0%. The Tier 2 teacher pass at Round 2 was pivotal for this dataset, producing the largest single-round gain across all experiments (mean accuracy +23.4 pp, from 28.3% to 51.7%). Recovery at Round 4 remained substantial (mean +23.3 %), showing that MMAD continued correcting errors in later rounds.

Answer stability at Round 3 was high for Agents P and Q (0.95 and 1.00 respectively), suggesting that the Tier 1 peer reasoning pass successfully locked in the gains established by the Tier 2 teacher at Round 2.

6.2 Commonsense Reasoning

CommonsenseQA presented a heterogeneous starting point across agents. Agent Q (Qwen2.5-7B)

Dataset	Agent	Accuracy (%)						Round 4	
		Base	R1	R2	R3	R4	R5	Drift%	Recov.%
GSM8K	M	5.0	15.0	25.0	45.0	75.0	80.0	0.0	30.0
	P	0.0	25.0	60.0	65.0	90.0	85.0	0.0	25.0
	Q	25.0	45.0	70.0	70.0	85.0	80.0	0.0	15.0
	Mean	*10.0*	*28.3*	*51.7*	*60.0*	***83.3***	*81.7*	***0.0***	*23.3*
CodeQA	M	47.6	57.1	61.9	66.7	66.7	61.9	9.5	9.5
	P	9.5	38.1	47.6	38.1	61.9	42.9	4.8	28.6
	Q	42.9	33.3	33.3	28.6	42.9	42.9	0.0	14.3
	Mean	*33.3*	*42.8*	*47.6*	*44.5*	***57.2***	*49.2*	*4.8*	*17.5*
CS1QA	M	28.0	50.0	57.0	55.0	53.0	50.0	9.0	7.0
	P	18.0	51.0	52.0	52.0	48.0	46.0	9.0	5.0
	Q	53.0	53.0	53.0	53.0	48.0	46.0	8.0	3.0
	Mean	*33.0*	*51.3*	*54.0*	*53.3*	***49.7***	*47.3*	*8.7*	*5.0*
CSQA	M	80.0	80.0	90.0	90.0	95.0	95.0	0.0	5.0
	P	0.0	60.0	90.0	75.0	90.0	80.0	0.0	15.0
	Q	95.0	85.0	100.0	95.0	85.0	90.0	15.0	5.0
	Mean	*58.3*	*75.0*	*93.3*	*86.7*	***90.0***	*88.3*	*5.0*	*8.3*

Table 3: Per-agent accuracy (%) at each MMAD round (Base = Phase 0, R1–R5 = Rounds 1–5), alongside Round 4 drift (correct-to-incorrect flip) and recovery (incorrect-to-correct flip) rates (%). Bold mean values indicate Round 4, the primary evaluation point. R5 is an additional Tier 1 round reported for completeness. Drift/Recovery rates are averaged across agents.

led at baseline with 95.0%, reflecting the domain's relative alignment with its pre-training; Agent M reached 80.0%; and Agent P began at 0.0%, suggesting a systematic failure mode for Phi-4-mini on this task format. MMAD delivered a mean gain of $+31.7$ pp by Round 4 (58.3% $\rightarrow$ 90.0%). Agent P's recovery was particularly striking: from 0% at baseline to 60% at Round 1 ($+60$ pp in a single Tier 1 peer-reasoning pass) and 90% at Round 2. This suggests that for CommonsenseQA, peer reasoning traces provided sufficient signal to largely correct Agent P's systematic baseline failures, even before teacher intervention. Agent Q exhibited the only notable drift event in this dataset: a drop from 100% at Round 2 to 95% at Round 3 (drift rate 5%) and then to 85% at Round 4 (drift rate 15%). This suggests a residual susceptibility to incorrect-majority pressure even after teacher guidance, though the absolute accuracy at Round 4 (85%) still represents a substantial improvement over Agent Q's baseline. The Tier 2 teacher pass at Round 2 achieved the peak mean ensemble accuracy of 93.3%, suggesting that for this dataset the teacher's guidance was most effective when applied before the peer-reasoning dynamics of Round 3 could introduce any backsliding.

6.3 CodeQA: Code Reading Comprehension

CodeQA showed steady but more modest improvement, with a mean gain of $+23.8$ pp by Round 4 (33.3% $\rightarrow$ 57.2%). Agent M improved most consistently, from 47.6% at baseline to 66.7% by Round 3, maintaining this level through Round 4. Agent P made strong gains from the extremely low baseline of 9.5%, reaching 61.9% at Round 4. Agent Q was relatively resistant to MMAD: after an initial drop from 42.9% at Round 1 to 28.6% in Round 3, it recovered to 42.9% in Round 4, matching but not surpassing its starting accuracy. The early drop in Agent Q's performance on CodeQA is notable. It represents an exception to the generally consistent accuracy gains observed across other tasks. We attribute this to lower peer answer quality, with Agents M and P performing poorly in early rounds, and Agent Q's higher baseline, making peer reasoning traces weaker and increasing sycophantic drift. The Round 4 teacher intervention halted the decline, restoring Agent Q to baseline accuracy. Round 4 drift rates on CodeQA were the highest: 9.5% for Agent M and 4.8% for Agent P, with a mean of 4.8%. While substantially lower than unstructured debate, these drift events indicate that CodeQA's short, precise answer requirements make maintaining correct answers under peer disagreement particularly challenging for some models.

6.4 CS1QA: Introductory Programming QA

CS1QA showed the smallest absolute gains ($+16.7$ pp at Round 4) and the most variable accuracy trajectory, reflecting the open-ended nature of

Dataset	Drift%	Recov.%	Stability%	Gain pp
GSM8K	0.0	23.3	76.7	+73.3
CSQA	5.0	8.3	86.7	+31.7
CodeQA	4.8	17.5	76.2	+23.8
CS1QA	8.7	5.0	69.7	+16.7

Table 4: Round 4 drift rate, recovery rate, stability, and accuracy gain relative to Phase 0 baseline, averaged across agents. Datasets sorted by accuracy gain.

System	Ph. 0	R1	R2	R3	R4	Gain
Conf-Debate	0.53	0.75	0.75	0.76	0.76	+0.23
MMAD	0.58	0.75	0.93	0.87	0.90	+**0.32**

Table 5: Mean accuracy on CommonsenseQA ($n = 500$ for Conf-Debate; $n = 20$ for MMAD). Both systems use Agents M, P, and Q. Conf-Debate augments standard peer debate with self-assessed confidence scores and has no teacher model. MMAD alternates peer-level MToM (odd rounds) and teacher-level MToM (even rounds).

its answers. Agents M and P improved from baseline (28% and 18%) to Round 2 peaks (57% and 52%) due to strong Tier 2 effects. Some of these gains eroded in later rounds, with mean Round 4 accuracy (49.7%) slightly below the Round 2 peak (54.0%). Agent Q, starting from a higher baseline of 53%, showed no improvement in Rounds 1–3 and and a slight decline to 48% at Round 4, indicating a ceiling effect and mild sycophantic drift.

This pattern suggests that for open-ended code explanation tasks, Tier 1 peer review at Round 3 can slightly destabilize gains from the Tier 2 teacher in Round 2, possibly because peer reasoning traces for subjective answers are more varied, offering weaker corrective signal than for tasks with unique correct answers. Drift rates at Round 4 averaged 8.7%, the highest across all datasets, corroborating this interpretation.

6.5 Round-by-Round Dynamics

Across all four datasets, a clear pattern emerges: Tier 1 (peer MToM) rounds produce initial accuracy improvements by revealing agents' reasoning and mitigating blind conformity to the majority. Tier 2 (teacher MToM) rounds reinforce these gains and resolve remaining shared misconceptions that peer debate alone cannot correct.

The tiered synergy is most evident on GSM8K: Tier 1 in Round 1 raises mean accuracy from 10% to 28.3%, creating diverse reasoning, and Tier 2 in Round 2 leverages this to reach 51.7%. Subsequent Rounds 3–4 further compound these gains, showing the benefit of a second alternating cycle.

Table 4 summarises drift and stability metrics at Round 4. A clear trade-off is visible: datasets with the largest accuracy gains (GSM8K, CSQA) also exhibit the lowest residual drift rates. This indicates that structured teacher ToM guidance is most effective on tasks with a unique correct answer. Open-ended tasks (CS1QA) exhibit higher drift and smaller gains, suggesting a need for richer or domain-specific teacher guidance.

6.6 Baseline Comparison

MMAD vs. Conf-Debate on CommonsenseQA. Tables 3 and 5 together reveal the key limitation of confidence weighting as a substitute for structured mutual awareness. Both systems reach the same mean accuracy at Round 1 (0.75), , confirming that a single round of peer exchange, whether using confidence scores or full reasoning traces, is similarly effective at aggregating complementary knowledge.

From Round 2 onward, however, the two systems diverge sharply: Conf-Debate plateaus entirely (0.75 at R1 to 0.76 at R4, a gain of less than 1 pp across three additional rounds), while MMAD continues to improve substantially, reaching 0.93 at Round 2 and 0.90 at Round 4—a +14.5 pp advantage over Conf-Debate at the final round. This plateau is due to miscalibrated self-reported confidence scores. At Round 4, Conf-Debate agents assign nearly identical mean confidence to incorrect answers (0.94–0.97) and correct ones (0.95–0.99), rendering the confidence signal largely ineffective for judging reasoning quality. Without a way to identify *why* an answer is wrong, peer pressure reinforces early consensus instead of correcting it. In contrast, MMAD's teacher uses teacher-level ToM to pinpoint each agent's reasoning errors and provide targeted guidance, yielding the large Round 2 jump (+18.3 pp) that Conf-Debate never achieves.

7 Discussion

Structured ToM outperforms confidence baseline. MMAD's two-tier MToM framework substantially improves accuracy across tasks and SLMs, outperforming confidence-based debate, with Round 4 improvements from +16.7 pp (CS1QA) to +73.3pp (GSM8K) over single-agent baselines. The Conf-Debate baseline, cannot reveal *where* an error occurs. An SLM observing a peer with confidence 0.9 has no basis for determining

whether that confidence is well-founded; an SLM reading a peer's full reasoning trace can identify a misread variable or arithmetic slip and appropriately discount the peer's conclusion. MMAD allows agents to detect missteps and properly weigh peer conclusions. This ability to leverage reasoning-trace evidence, rather than scalar uncertainty, underlies MMAD's performance advantage.

The teacher's role is indispensable for hard tasks. GSM8K, the hardest baseline task for these SLMs, shows the largest improvement and the clearest elimination of drift. The single-agent baseline accuracy averages just 10%: agents collectively lacked the knowledge to resolve disagreements through peer debate alone. The Tier 2 teacher intervention at Round 2 accounts for the largest single-round accuracy jump observed in our experiments (28.3% to 51.7%), by identifying the specific arithmetic or setup error in each agent's reasoning and redirecting attention without revealing the answer. This Socratic guidance pattern (Zeng et al., 2022), which asks questions instead of giving answers, is crucial: simply providing the answer would boost accuracy but not teach agents to assess peer reasoning in later rounds.

Agent heterogeneity affects MMAD dynamics. Agent P (Phi-4-mini) consistently started from the lowest baseline but achieved the largest round-4 improvements for all datasets. We attribute this to two factors: i) a high improvement ceiling and ii) Phi-4-mini's responsiveness to structured reasoning traces, which better leverages its instruction-following and in-context reasoning than single-turn inference. Agent Q (Qwen2.5-7B) showed the highest baselines but also the greatest susceptibility to sycophantic drag in certain rounds, most visibly on CSQA (Round 4 drift 15%) and CodeQA (Rounds 1–3 accuracy decline). This pattern suggests that high-baseline agents may benefit from stronger stability reinforcement in the Tier 1 prompt, such as emphasizing "maintain your answer if peers cannot identify a specific error".

Open-ended tasks present a harder challenge for ToM. CS1QA, which features open-ended explanations rather than unique correct answers, shows consistently higher drift rates and smaller gains than the other three datasets. In tasks with unique answers, the teacher can precisely localise an error ("your loop condition evaluates to True when the counter reaches N, not $N-1$"). For open-ended explanations, the boundary between an acceptable and unacceptable answer is fuzzier, making precise error localisation harder and the teacher's teacher-level ToM guidance less targeted. Future work might address this by providing the teacher with rubric-style evaluation criteria in addition to the gold answer.

8 Conclusion

We present MMAD, a two-tier MToM debate framework that eliminates sycophantic drift in small language model debate. Our experiments show that SLMs in standard peer debate often abandon correct answers under majority pressure—not due to superior peer reasoning, but because they treat consensus as a proxy for correctness without a way to evaluate reasoning quality.

MMAD intervenes on two complementary levels. Tier 1 requires agents to read and critically assess peer reasoning before updating their answers, introducing first-level Theory of Mind and breaking naive majority-following. Tier 2 employs a 30B teacher model that applies teacher-level ToM across all agents, providing personalized Socratic guidance that identifies each agent's specific errors without revealing the answer, producing the largest gains on difficult tasks where peer debate alone falls short.

Our experiments on four datasets in mathematics, code, and commonsense reasoning show substantial Round 4 gains over single-agent baselines: $+73.3$ pp on GSM8K, $+31.7$ pp on CommonsenseQA, $+23.8$ pp on CodeQA, and $+16.7$ pp on CS1QA. On GSM8K, sycophantic drift drops to 0%, showing that structured mutual awareness can fully eliminate this failure on tasks with unique answers. These gains are achieved without any fine-tuning, making MMAD directly applicable to any instruction-tuned SLM.

The results also reveal directions for future work: personalised stability reinforcement for high-baseline agents susceptible to sycophantic drag; rubric-augmented teacher prompting for open-ended tasks; and distillation of the teacher's ToM guidance capability into a smaller, more deployment-efficient model. More broadly, our findings suggest that SLM collaboration is limited less by knowledge—correct reasoning is usually present—than by the ToM ability to judge which reasoning to trust. Structured mutual awareness, via Theory-of-Mind prompting, provides a principled way to develop this capacity.

Limitations

Our evaluation is subject to several limitations. First, our teacher model (Qwen3-Coder-30B-A3B-Instruct) introduces additional inference cost at even rounds; while the complete pipeline runs on a single consumer-grade GPU server, real-time deployment at scale may require distilling the teacher's ToM guidance capability into a smaller model. Second, while we evaluate across four datasets and three SLM families, our results may not generalise to all task types (e.g., long-form generation or structured prediction). Third, the confidence scores in Conf-Debate are self-reported and likely miscalibrated; a stronger baseline with externally calibrated confidence might narrow the gap with MMAD, though it could not provide the reasoning-trace transparency that underlies the key gains. Fourth, we use greedy decoding throughout for reproducibility; stochastic sampling might reveal additional variance in sycophantic drift patterns.

Ethical considerations

The experiments in this work involve prompting instruction-tuned small language models (Mistral-7B-Instruct-v0.3, Phi-4-mini-instruct, Qwen2.5-7B-Instruct) and a larger teacher model (Qwen3-Coder-30B-A3B-Instruct) as agents in a multi-agent debate pipeline. These models are used solely as research subjects to measure reasoning accuracy, sycophantic drift, and recovery rates; they are not deployed in any downstream application. No personal data is involved: all evaluation datasets (GSM8K, CodeQA, CS1QA, CommonsenseQA) are publicly released research benchmarks. The research in this paper was assisted by AI tools for grammar checking and code formatting; all scientific content, experimental design, analysis, and conclusions are the sole work of the authors.

Acknowledgements

[To be filled in for the final version. Anonymised for review.]

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awasthi, Ammar Ahmad Awan, Nguyen Bach, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic ai. *Preprint*, arXiv:2506.02153.

Jing Bi, Yuting Wu, and Weiwei Xing. 2025. Enhancing the reasoning capabilities of small language models via solution guidance fine-tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9074–9084, Abu Dhabi, UAE. Association for Computational Linguistics.

Justin Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2025a. MAgICoRe: Multi-agent, iterative, coarse-to-fine refinement for reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32663–32686, Suzhou, China. Association for Computational Linguistics.

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

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025b. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11850–11881. PMLR.

Wei Fan, JinYi Yoon, and Bo Ji. 2026. imad: Intelligent multi-agent debate for efficient and accurate llm inference. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)*. Oral Presentation.

Tao Feng, Yicheng Li, Chenglin Li, Hao Chen, Fei Yu, and Yin Zhang. 2024. Teaching small language models reasoning through counterfactual distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842, Miami, Florida, USA. Association for Computational Linguistics.

X. Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony G. Cohn, and Michael J. Wooldridge. 2024. A notion of complexity for theory of mind via discrete world models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2964–2983, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Avneet Kaur. 2025. Echoes of agreement: Argument driven sycophancy in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22803–22812, Suzhou, China. Association for Computational Linguistics.

Sung Won Kim and Daniel Khashabi. 2025. Challenging the evaluator: LLM sycophancy under user rebuttal. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22461–22478, Suzhou, China. Association for Computational Linguistics.

Changyoon Lee, Yeon Seonwoo Moon, and Alice Cho. 2022. CS1QA: A dataset for assisting code-based question answering in an introductory programming course. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–197, Singapore. Association for Computational Linguistics.

Xiang Li, Shizhu He, Fangyu Lei, Jun Yang, Tianhuang Su, Kang Liu, and Jun Zhao. 2024a. Teaching small language models to reason for knowledge-intensive multi-hop question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7804–7816, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Chenxiao Liu and Xiaojun Wan. 2021. CodeQA: A question answering dataset for source code comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2584–2598, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lili Pan, Lei Li, and William Wang. 2024. Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. 2025. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22112–22133, Vienna, Austria. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10228–10241, Abu Dhabi, UAE. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Eitan Wagner, Nitay Alon, Joseph M. Barnby, and Omri Abend. 2025. Mind your theory: Theory of mind goes deeper than reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26658–26668, Vienna, Austria. Association for Computational Linguistics.

Qiaosi Wang and Ashok K Goel. 2022. Mutual theory of mind for human-ai communication. In *IJCAI Workshop on Communication in Human-AI Interaction (CHAI)*.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

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

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

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *Preprint*, arXiv:2204.00598.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*.