When Debate Messages Help or Hurt: Diagnosing Communication in Self-Distillation

Jiatao Yao*
iFLYTEK Co., Ltd.
Hefei, China
tritonyao@163.com

Bovin Tan*
Guanghua Law School
Zhejiang University
Hangzhou, China
bovintan@zju.edu.cn

Abstract

Multi-agent debate is widely used to enhance reasoning in language models, yet the exchanged messages are rarely analyzed. We study these messages in a Debate+SFT self-distillation loop by tracking their round-level evolution, intervening on peer messages with T-CIC at inference time, and comparing against matched Solo controls. Across three 5-seed discovery tasks–MATH, GSM-Hard, and 2WikiMultiHopQA–we observe weak refinement on MATH (+0.9pp for Debate versus +9.0pp for Solo), harmful homogenization under the default temperature schedule on GSM-Hard (-11.3pp, with messages 27.5% longer and more self-similar), and mild inflation on 2WikiMHQA (-0.5pp with moderate length growth). A 2-seed T-CIC check on round-4 adapters shows different forms of message dependence: MATH collapses when peer messages are removed, shuffled, or replaced (-16 to -23pp), GSM-Hard is selectively sensitive to mismatched context, and 2WikiMHQA is mostly stable except under cross-problem shuffling. Debate messages should therefore be measured and stress-tested before they are trusted as a training signal.

1 Introduction

Multi-agent debate is a prevalent mechanism for improving LLM reasoning, premised on the idea that agents can catch mistakes made by one another (Du et al., 2024; Liang et al., 2024; Subramaniam et al., 2025). When integrated into self-distillation (Debate+SFT), this approach appears attractive on two fronts: the model trains on its own outputs, while the debate process ostensibly refines the reasoning traces (Hosseini et al., 2024). However, this integration introduces a subtle yet critical risk. The training signal does not consist solely of the final answer, but also encompasses the natural-language messages exchanged during the debate. Although these messages are often assumed to serve as useful critiques, they may instead render subsequent training data more repetitive or introduce misleading supervision.

Prior work on debate often evaluates the final group answer while the exchanged text itself remains largely unexamined (Du et al., 2024; Liang et al., 2024; Subramaniam et al., 2025). This approach is insufficient for self-distillation loops, where messages play a dual role: they influence the answer of the current round, and filtered versions of them serve as supervision for the subsequent round. If the messages become degenerate, for example, if agents merely reproduce the errors of their peers, the training loop will reinforce this failure mode, a phenomenon that endpoint accuracy alone cannot detect.

We argue that debate communication should be treated as a conditional resource rather than as free supervision. Messages are beneficial only when they support correct answers or provide reliable training signals, and are harmful when their evolution is tied to worse answers or to homogenization. This operational view implies that longer, more diverse, or more confident messages are not intrinsically advantageous. Their value must be assessed empirically within the training process.

To measure this conditional resource, we introduce a diagnostic framework consisting of three components. First, we track round-level message signals (length, lexical diversity, self-similarity) to detect trajectory anomalies. Second, we introduce **Task-level Communication Intervention** (T-CIC), an inference-time stress test that alters peer messages without retraining to determine whether final answers genuinely depend on message content. Third, we compare Debate+SFT against matched Solo controls and ablations to isolate communication effects from data retention or persona wording.

We applied this diagnostic across a 5-seed experiment, which reveals marked variability, proving

*Both authors are corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 18461–18473
October 24-29, 2026 ©2026 Association for Computational Linguistics

that debate messages must be measured and stress-tested before they are trusted as a training signal. The same Debate+SFT loop yields three distinct regimes depending on the task. It produces weak refinement on MATH (+0.9pp), harmful homogenization on GSM-Hard (−11.3pp, with longer and more self-similar messages), and mild inflation on 2WikiMHQA (−0.5pp). These labels describe observed trajectories rather than fixed task categories.

Contributions.

- We propose a diagnostic framework that formalizes communication quality as an empirical property of the Debate+SFT loop, tracking message trajectories (length, diversity, homogenization) in parallel with accuracy.
- We introduce T-CIC, an inference-time stress test that tests the dependence of final answers on peer messages without retraining, thereby distinguishing task-aware content dependence from surface-form sensitivity.
- We demonstrate that debate messages constitute a conditional resource rather than free supervision: the same loop yields weak refinement, harmful homogenization, or mild inflation across tasks, challenging the assumption that multi-agent exchange universally improves self-distillation.

The rest of the paper defines the diagnostics (Section 3), applies them to the six-task suite (Sections 4–5), and discusses how the resulting diagnostics should be interpreted and used (Section 6).

2 Related Work

2.1 Multi-Agent LLM Debate

Multi-agent debate is motivated by a simple idea: agents may catch one another's mistakes when they exchange reasoning traces (Du et al., 2024; Liang et al., 2024). Later work studies variants such as confidence-weighted consensus (Chen et al., 2024) and co-evolving agent interactions (Xue et al., 2026). Closest to our setting, Subramaniam et al. (2025) combine debate with SFT over multiple rounds, but prior work typically evaluates debate through endpoint accuracy or agent specialization. Our focus is different: we diagnose the messages inside Debate+SFT and ask when they behave like useful context versus harmful training input.

2.2 Self-Distillation

Our loop belongs to the self-training and self-distillation family. STaR (Zelikman et al., 2022) bootstraps on correct rationales, ReST^{EM} (Singh et al., 2024) generalizes the grow–improve pattern, and V-STaR (Hosseini et al., 2024) uses fresh adapters with conservative filtering. OPTIMA (Chen et al., 2025) explicitly rewards token efficiency; by contrast, we do not optimize message form, so compression or inflation is an observed outcome of the debate loop rather than a training objective.

2.3 Model Collapse

Recursive training on model-generated data can narrow the learned distribution over time (Shumailov et al., 2024; Alemohammad et al., 2024). We study a related but distinct phenomenon: the distribution of *inter-agent messages* changes inside a vote-filtered debate loop. The Debate-vs-Solo comparison helps separate ordinary self-distillation drift from changes that come from the communication channel: compression, diversity shifts, and homogenization across agents.

2.4 Causal Communication

Emergent-communication work asks when messages affect other agents' actions, often through controlled signaling games and CIC-style interventions (Chaabouni et al., 2019; Lazaridou et al., 2017; Havrylov and Titov, 2017; Eccles et al., 2019; Lowe et al., 2019). LLM debate differs because agents already have a strong language prior and are updated by SFT, not policy gradient. We therefore frame the phenomenon as induced communication and use T-CIC (§3.4) as a task-level stress test: final predictions should be stable under paraphrase, but should change when useful peer messages are removed or mismatched.

3 Method

Our goal is to make the messages in Debate+SFT inspectable. For each run, we ask three questions. How do messages change over rounds? Do final answers still work if the messages are paraphrased, removed, or replaced? And do the observed changes look like useful refinement, misleading homogenization, or mostly harmless inflation? The Debate+SFT loop is the setting; the contribution is the diagnostic view of the messages inside it. Full pseudocode, implementation details, and the pre-

registered classification rule are in Appendices A.1, A.2, and C.3.

Concretely, we combine round-level message signals (length, MATTR, Self-BLEU, and unanimous rate), the T-CIC stress test, and matched Solo/ablation controls. The diagnosis does not prescribe a new training algorithm. It tells us whether messages behave like useful context, misleading guidance, or extra text that the model largely ignores.

3.1 Pipeline Overview

The implementation keeps the iterative training loop separate from the diagnostic analysis. During training, each round evaluates the current adapter on a fixed evaluation split, resamples the training split with a round-specific seed, saves both train and eval transcripts, applies the configured keep rule to the training transcripts, and fits the next LoRA adapter. The logged message metrics are used for analysis, not for deciding which examples enter SFT. In our main runs the keep rule is unanimous-correct for Debate and correctness for Solo. For the final reported round, separate scripts read the saved round-4 checkpoint and outputs and run T-CIC on that adapter. Thus T-CIC is a post-hoc inference-time validation: it perturbs the messages shown before the final answer, changes no weights, and never feeds back into sample selection.

3.2 Debate + Self-Distillation Loop

Debate protocol. For each problem x, Debate uses $N{=}3$ agents with the same base model and different personas: careful calculator, skeptical critic, and symbolic reasoner. At turn 1, agents independently generate solutions. At turn 2, each agent sees the other agents' turn-1 messages (self-exclusion), critiques them, and emits a final answer. We use temperature 0.7 for turn 1 and 0.3 for turn 2, following prior debate work (Du et al., 2024); the analysis tests the swapped schedule.

Evaluation vote vs. training filter. We use *majority vote* only to compute the Debate group prediction: $\hat{a}$ is correct when at least two of three final answers match the gold answer. We use a stricter *unanimous-correct* filter for training: a Debate transcript is retained for SFT only when all three agents agree and that shared answer is correct. For Solo ($N{=}1$), the same rule reduces to a plain correctness filter. Throughout the paper, accuracy refers to the evaluation vote, while self-distillation data refers to unanimous-correct training transcripts.

SFT update. Each retained Debate transcript contributes one randomly sampled turn-2 message as the training target, preventing Debate from contributing three times as many rows as Solo. Training resamples 500 candidate training problems per round but keeps the 200-problem evaluation split fixed within a cell, so round-to-round changes are not caused by changing eval items. Each round uses the current adapter to collect transcripts, then trains a new LoRA adapter from the base model on that round's retained targets, following STaR, ReST$^{\text{EM}}$, and V-STaR-style self-distillation (Zelikman et al., 2022; Singh et al., 2024; Hosseini et al., 2024). We run $R{=}5$ rounds; Appendix A.1 gives the complete loop and Appendix C.3 gives the hyperparameters.

3.3 Post-Hoc Message Metrics

After training, each round's 200-problem evaluation split is used for diagnosis. We measure turn-2 messages because they are the communicated content used for final answers. We report:

- **Accuracy**: majority-vote accuracy for Debate and single-answer accuracy for Solo.
- **Mean length**: average token count under the Qwen2.5 tokenizer.
- **MATTR**$_{50}$: moving-average type-token ratio (Covington and McFall, 2010), used instead of raw TTR to reduce length confounds.
- **Self-BLEU**: mean pairwise BLEU-4 among the three agents' turn-2 messages on the same problem; higher values indicate homogenization.
- **Unanimous rate** and $n_{\textbf{train}}$: the fraction and count of examples passing the unanimous-correct filter.

3.4 T-CIC: Task-Level Communication Intervention

Round-level metrics tell us that messages change; T-CIC asks whether those messages matter for the final answer. T-CIC is run after the round-4 adapter has been selected for final reporting. We do not retrain anything, resample training data, or update the adapter. Instead, the same round-4 adapter answers the same questions while seeing altered peer messages before the final answer. Following the

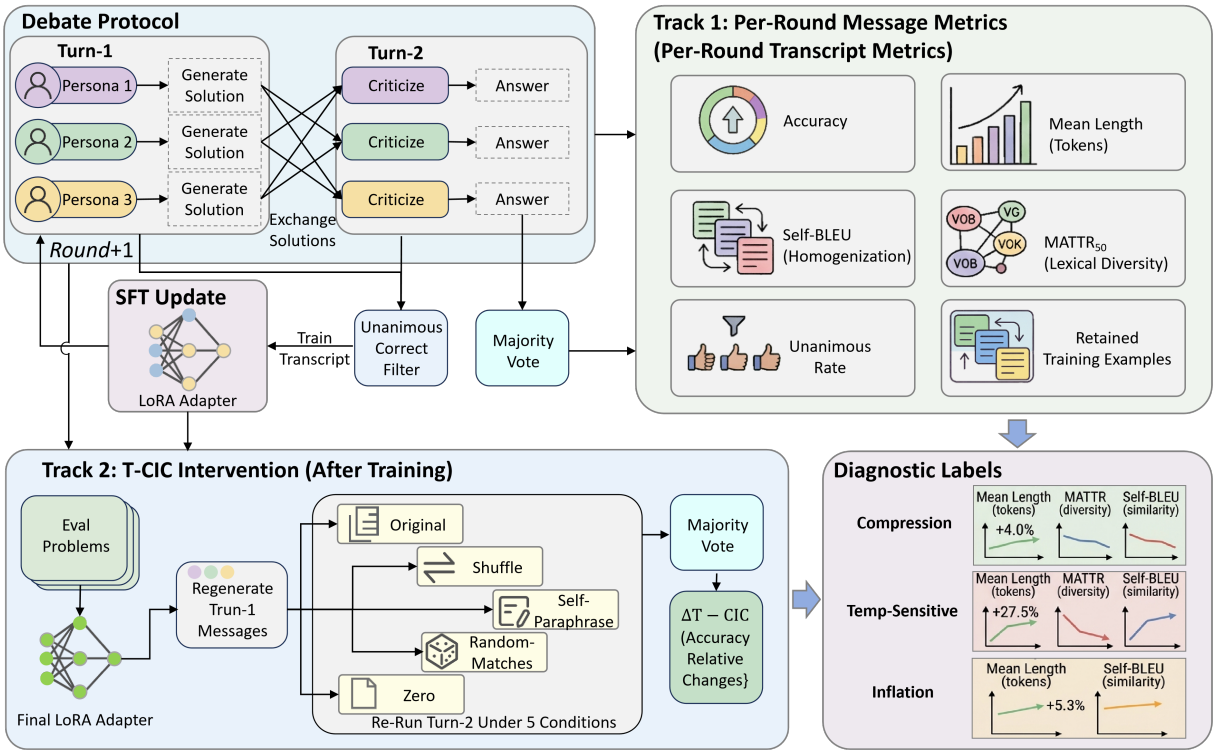


Figure 1: **End-to-end experimental pipeline.** Each training round runs the current model or adapter to collect transcripts, logs evaluation metrics for analysis, filters only the training transcripts, and trains the next LoRA adapter. T-CIC is executed post hoc on the round-4 adapter, so it validates message sensitivity without changing sample selection or model parameters.

idea that communication can be probed by intervening on messages (Lowe et al., 2019), we treat T-CIC as an inference-time stress test for natural-language debate, not as a full causal estimate of training dynamics.

For each problem, the round-4 adapter first regenerates turn 1 messages for the same evaluation examples. We then re-run turn 2 under five conditions:

- **Original**: the unmodified turn-1 summaries generated by the round-4 adapter.
- **Self-paraphrase**: same content, rewritten by the same checkpoint.
- **Zero**: an explicit empty-message placeholder.
- **Shuffle**: peer messages from other examples in the same task split.
- **Random-matched**: unrelated text matched to the original tokenizer length.

We report $\Delta_{\text{T-CIC}} = \text{Acc}_{\text{intervention}} - \text{Acc}_{\text{Original}}$. Negative values mean the intervention hurts relative to the original messages. We interpret the five accuracies jointly. A paraphrase drop suggests surface-form sensitivity. A Zero drop suggests that peer content is useful. A Shuffle or Random-matched drop suggests that the useful content is tied to the problem, rather than to message length alone. Unlike a training ablation, T-CIC changes no data, weights, or task distribution; it isolates inference-time sensitivity to the messages shown in turn 2.

3.5 Controls and Analysis

The central comparison is **Debate** (N=3) **vs. Solo** (N=1) with matched base model, rounds, seeds, evaluation split, and filtering principle (unanimous-correct for Debate; correctness for Solo). Three ablations address the main confounds:

(A) Matched-train-size Solo downsampled to Debate's retained-data floor, testing whether Debate degradation is just data starvation.

(B) Identical-persona Debate with a shared generic persona and distinct seeds, testing whether differentiated persona wording drives the effect.

(C) Llama-3.1-8B A cross-model check on the discovery tasks, testing whether the main diagnostic pattern is Qwen-specific.

We assign diagnostic labels with the pre-registered decision rule in Appendix A.2: labels depend on joint slopes of accuracy, length, and lexical diversity rather than on endpoint accuracy alone. In brief, harmful labels require degraded accuracy together with deteriorating message signals, while inflation labels require significant length growth without a matching accuracy gain. We use these labels as compact summaries of observed trajectories, not as a fixed taxonomy of tasks. Main effects are estimated with mixed-effects models and Benjamini–Hochberg correction over the pre-registered metric–task grid; the main text highlights the effect sizes most relevant to the diagnostic claims, with the full decision rule retained in the appendix.

4 Experiments

4.1 Experimental Setup

We evaluate whether debate messages help or hurt across tasks that require different kinds of reasoning. Three tasks form the main 5-seed discovery set and support the full trajectory analysis. The T-CIC validation is run on round-4 adapters for 2 seeds per discovery task. Three additional tasks serve as 2-seed related-task checks under the same debate and training pipeline; we use them to probe recurrence, not as a second discovery set.

Model and training. Qwen2.5-7B (bf16) with fresh-from-base LoRA (r=16, α=32, lr 2×10^{-4}, 2 epochs) per round (Hosseini et al., 2024). Inference uses vLLM v1 with temperature 0.7 (turn 1) / 0.3 (turn 2).

Tasks. The discovery set contains **GSM-Hard** (numeric arithmetic), **MATH** L3–4 (symbolic LaTeX), and **2WikiMultiHopQA** (multi-hop QA). The related-task set contains **HumanEval** (code synthesis), **StrategyQA** (strategic reasoning), and **CommonsenseQA** (commonsense multiple-choice QA). For each discovery cell, we keep a fixed 200-problem evaluation split and resample 500 candidate training problems each round; related-task runs use 2 seeds and check whether the same patterns appear beyond the main tasks.

Arms and seeds. The discovery grid contains 3 tasks $\times$ 2 arms $\times$ 5 seeds $=$ 30 main training cells. The validation grid adds 3 related tasks $\times$ 2 arms $\times$ 2 seeds $=$ 12 validation cells. **Debate** (N=3 agents, differentiated personas, unanimous-correct training filter) is compared against **Solo** (single agent, correctness filter). Discovery seeds: {17, 42, 2024, 1337, 7}; validation seeds: {17, 42}. R=5 rounds per cell.

Metrics. We track final accuracy, message length, MATTR_{50} (Covington and McFall, 2010), and Self-BLEU-4. We also report unanimous vote rate and the number of retained training examples, since communication can matter either by changing answers or by changing what enters the next round of training.

4.2 Main Results

Table 1 summarizes the 5-seed discovery set. The result is not "debate helps" or "debate hurts." The same loop behaves differently by task. On MATH, Debate improves slightly but less than Solo. On GSM-Hard, Debate is clearly harmful: accuracy drops while messages get longer and more similar. On 2WikiMHQA, accuracy changes little, but messages still show mild inflation. The labels in Table 1 are reader-facing summaries of the formal cell-level rule and matched-Solo effect sizes; the rule and the relationship between the two naming levels are given in Appendix A.2. Appendix C.2 reports related-task checks at a lower evidence tier. We next ask whether the saved messages affect final answers at inference time.

Weak-refinement family. On MATH, Debate improves only slightly from R0 to R4 (+0.9pp), while Solo improves much more (+9.0pp). Debate messages grow mildly from 282.2 to 293.6 tokens (+4.0%), MATTR falls from .617 to .563, and Self-BLEU falls from .622 to .588. We read MATH as weak refinement: communication is not collapsing, but it also does not explain the stronger gains seen in Solo.

Harmful-homogenization family. Under the default schedule (0.7 gen / 0.3 crit), GSM-Hard shows the clearest harmful communication pattern. Debate accuracy declines by 11.3pp, while Solo is essentially flat (+0.4pp). Debate length rises from 299.2 to 381.4 tokens (+27.5%), MATTR falls from .626 to .407, and Self-BLEU rises from .597 to .714. The failure is therefore not just verbosity: the messages become longer, less lexically diverse, and more mutually similar while accuracy falls. The related StrategyQA run shows a similar sensitivity to the generation and critique schedule (Appendix C.2).

Task	Debate ΔAcc	95% CI	Solo ΔAcc	Message signal	Diagnostic regime
MATH	+0.9pp	$[-6.5, +8.3]$	+9.0pp	Len +4.0%, MATTR −.054, S-BLEU ↓	Weak refinement
GSM-Hard	−11.3pp	$[-19.4, -3.2]$	+0.4pp	Len +27.5%, MATTR −.219, S-BLEU ↑	Harmful homogenization
2WikiMHQA	−0.5pp	$[-3.3, +2.3]$	+4.2pp	Len +5.3%, S-BLEU ↑	Mild inflation

Table 1: **Discovery-set diagnostic summary.** R0→R4 changes for Debate+SFT and matched Solo on the 5-seed discovery tasks. Positive ΔAcc means final-round accuracy improves over round 0; the 95% CI column reports intervals over the five seed-level Debate changes; the message-signal column records the main Debate communication pattern.

Mild-inflation family. On 2WikiMHQA, Debate accuracy is nearly flat but slightly negative (-0.5pp), while Solo improves by 4.2pp. Debate messages lengthen from 249.9 to 263.1 tokens (+5.3%), and Self-BLEU rises from .437 to .453. This is mild inflation rather than a collapse: the messages expand and become more alike, but the accuracy effect is small.

5 Analysis

5.1 Discovery-Set Trajectories

Figure 2 shows why we look beyond final accuracy. GSM-Hard does not merely end lower; it moves downward as messages become longer, less lexically diverse, and more self-similar. MATH and 2WikiMHQA have weaker endpoint effects, so we describe them more cautiously as weak refinement and mild inflation rather than as strong monotone trends.

5.2 T-CIC: Task-Level Communication Intervention

Message trajectories show that communication changes, but they do not prove that the final answer uses the messages. T-CIC tests this directly. It keeps the adapter, question, and decoding fixed, and changes only the peer messages shown before the final answer. **Original** uses the round-4 adapter's regenerated turn-1 messages; **Self-paraphrase** rewrites them; **Zero** removes them; **Shuffle** swaps messages across problems; and **Random-matched** uses length-matched random text. For each intervention, we re-run turn 2 for all three agents and score the majority-vote final answer. Results average over 200 eval problems $\times$ 2 seeds. We report $\Delta_{\text{T-CIC}} = \text{Acc}_{\text{intervention}} - \text{Acc}_{\text{Original}}$, so negative values mean that the intervention made final answers worse than the original debate messages.

Table 2 shows three different inference-time sensitivities to the saved messages. MATH is the most message-dependent cell: removing, shuffling, or replacing messages drops accuracy by 16–23pp. GSM-Hard is selectively sensitive: shuffled and random-matched context hurt by 13–19pp, while deleting the messages is close to Original (-2.5pp), consistent with a channel that mixes useful cues with unreliable ones. 2WikiMHQA is the boundary case: paraphrases, removal, and random-matched text remain near Original, but shuffling across problems still hurts by 6.0pp. Thus T-CIC should be read as a task-aware message-sensitivity test, not as a claim that all peer context is uniformly helpful or harmful.

5.3 Controls and Boundary Conditions

Two controlled checks rule out simple explanations for the GSM-Hard result. First, the decline is not only a retained-data problem: even when Solo is downsampled to the Debate data floor, it remains above the corresponding Debate baseline. Second, persona wording matters but does not fully explain the effect: identical-persona Debate improves accuracy while lowering MATTR.

Condition	R4 Acc	R4 MATTR	n_{train}
Debate baseline	.378	.690	113
Debate identical-p	.543	.653	241
Solo baseline	.492	.618	242
Solo matched	.510	.591	62

Table 3: GSM-Hard controls on the ablation subset. Confidence intervals and full details are in Appendix B.1.

The temperature schedule is another boundary condition. Under the default 0.7/0.3 generation/critique schedule, GSM-Hard Debate drops by 5.7pp in the ablation subset; under the swapped 0.3/0.7 schedule, it instead rises by 3.5pp (Appendix B.2). We therefore treat the GSM-Hard failure mode as schedule-sensitive rather than as a universal debate failure. Compact qualitative examples are placed at the end of the appendix for readers who want to inspect the message channel directly (Appendix D).

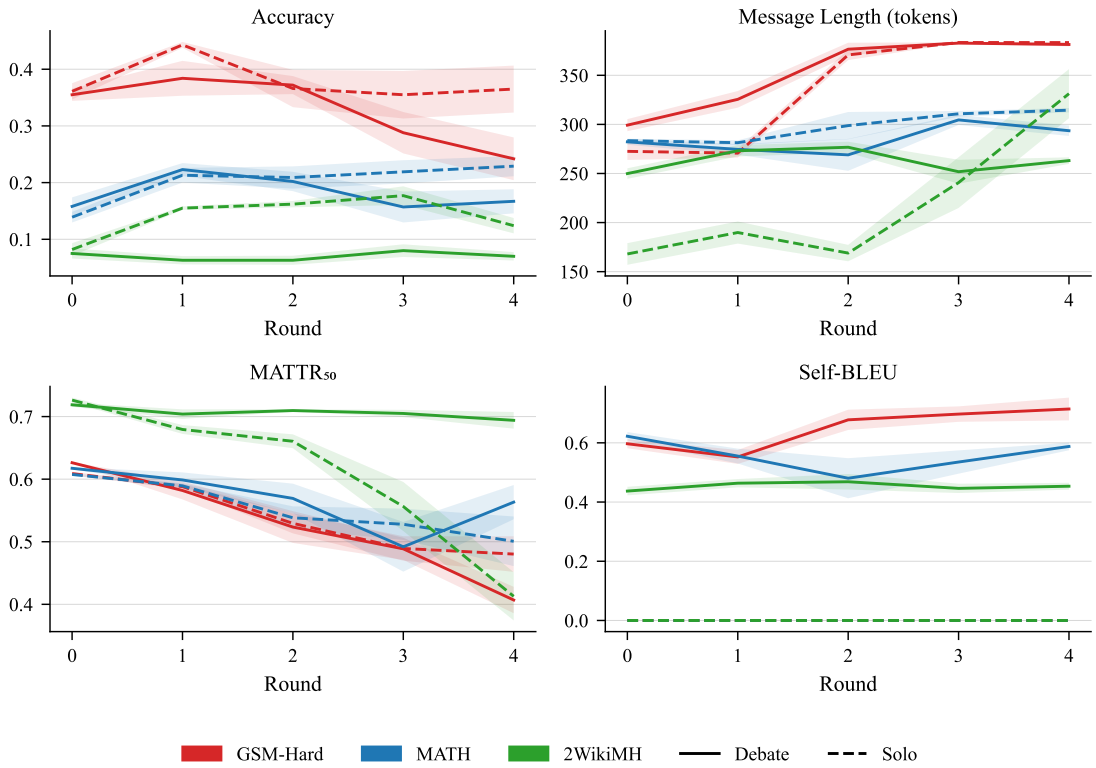


Figure 2: Round trajectories for the three 5-seed discovery tasks. The regimes in Table 1 arise from trajectory-level changes rather than isolated endpoints.

Intervention	GSM-Hard	MATH	2WikiMHQA
Original	.255	.270	.178
Self-paraphrase	.203 (−.053)	.195 (−.075)	.160 (−.018)
Zero	.230 (−.025)	.108 (−.163)	.193 (+.015)
Shuffle	.070 (−.185)	.058 (−.213)	.118 (−.060)
Random-matched	.123 (−.133)	.038 (−.233)	.185 (+.008)

Table 2: T-CIC validation on round-4 Debate adapters. Original reports accuracy; parentheses report $\Delta_{\text{T-CIC}}$ relative to Original. Each task uses 2 seeds and 400 total examples.

6 Discussion

6.1 Communication as a Conditional Resource

The practical lesson is that debate messages are not free supervision. They can shape both the final answer and the examples that survive into the next round. When messages become longer and more self-similar while accuracy falls, as on GSM-Hard, the loop is not simply collecting better rationales; it is reinforcing a particular communication style. T-CIC also rules out a simple fix. Deleting messages can severely hurt MATH, mildly hurt GSM-Hard, and leave 2WikiMHQA near its original accuracy. Communication should therefore be measured and stress-tested before it is trusted as either inference context or training signal.

6.2 Practical Diagnostics

The framework suggests three practical checks for iterative multi-agent training and inference.

- **Track length and Self-BLEU jointly.** Rising length together with rising Self-BLEU is the clearest warning sign of harmful inflation.
- **Compare MATTR and Self-BLEU against accuracy.** Falling MATTR with rising Self-BLEU and falling accuracy signals homogenized communication rather than genuine reasoning improvement.
- **Watch the unanimous-correct pass rate.** A low pass rate signals consensus breakdown and warns that the debate channel is becoming unreliable as a source of retained training signal.

Task	Debate	Solo	Fallback	vs. Debate
MATH	.167	.229	.188	+.021
GSM-Hard	.242	.365	.365	+.123
2WikiMHQA	.070	.124	.114	+.044
Macro	.160	.239	.222	+.063

Table 4: Diagnosis-guided fallback using early communication signals. The policy improves over always-Debate but does not yet beat always-Solo.

These checks are not yet a full intervention policy, but they make the failure mode actionable. As a post-hoc utility check, we let a simple policy observe Debate through round 2 and fall back to Solo at round 4 when at least three warning signals fire: rising length, rising Self-BLEU, falling MATTR, and falling unanimous rate. This does not retrain models or use round-4 accuracy to choose the arm, but it tests whether the diagnosis can guide a useful decision.

Table 4 improves macro accuracy over always-Debate by 6.3pp, mainly by avoiding the GSM-Hard failure mode. It also shows the limitation: always-Solo still wins on this suite, so the current result should be read as evidence that the diagnostics are decision-relevant, not as a final training policy.

6.3 Preliminary Recurrence Beyond the Discovery Set

The related-task runs are not meant to define a stable taxonomy, but they show that similar symptoms recur outside the discovery set. HumanEval produces shorter outputs and about +10pp accuracy, StrategyQA loses about 4.3pp under the default temperature schedule, and CommonsenseQA lengthens outputs by about 22% while Self-BLEU rises from about 0.42 to 0.58. Because these are 2-seed pilot checks, they support the discovery-set reading only at the family level: communication behavior is task-dependent, and the failure mode must be diagnosed rather than assumed from accuracy alone.

7 Conclusion

This paper shows that communication in debate self-distillation is not automatically beneficial. By tracking message trajectories and intervening on the messages with T-CIC, we can identify when peer messages provide useful context, when they become unreliable, and when they mostly add surface variation. Across the three 5-seed discovery tasks, the same Debate+SFT loop yields weak refinement, harmful homogenization, and mild inflation; three 2-seed related-task pilots and cross-model checks on Llama provide lower-tier support for the same diagnostic view. The broader implication is straightforward: multi-agent self-distillation should treat communication as something to measure and stress-test, rather than obscure it behind aggregate accuracy.

8 Limitations

Scale. Main results use Qwen2.5-7B; 70B+ effects are unverified. **Training.** LoRA SFT only; RLHF may yield qualitatively different dynamics. **Rounds.** R=5; longer loops may exhibit new regimes. **Tasks.** The discovery set contains three reasoning tasks, and the validation set adds three related tasks; open-ended generation remains untested. **Evidence tiers.** Five-seed discovery cells support the main claims; smaller validation, ablation, T-CIC, and cross-model cells are used as directional checks and boundary tests. **Implementation.** The identical-persona ablation corrects an earlier bug; corrected accuracy (.543) differs from the buggy run ($\sim$.398). **Cross-model.** The Llama-3.1-8B check (Appendix C.1) is preliminary and used as directional support rather than as a main claim. GSM-Hard and 2WikiMHQA show Debate trailing matched Solo at round 4; MATH remains too small to support a strong cross-model conclusion. **T-CIC coverage.** The final task-aware T-CIC protocol is complete for the three discovery tasks; it tests inference-time message sensitivity rather than full retraining dynamics. **Method scope.** The paper offers diagnostics rather than an optimized communication-aware training algorithm; policies built on top of these diagnostics remain future work.

Ethics Statement

This work involves no human subjects. All models (Qwen2.5-7B and Llama-3.1-8B) and datasets (GSM-Hard, MATH, 2WikiMultiHopQA, HumanEval, StrategyQA, CommonsenseQA) are publicly available. Code and message logs are available at https://github.com/THEKUNzz/multi-agent-may-hurt.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2024. Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*, pages 53581–53608.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6290–6300.

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.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2025. OPTIMA: Optimizing effectiveness and efficiency for LLM-based multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11534–11557, Vienna, Austria. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

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 11733–11763. PMLR.

Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, and Thore Graepel. 2019. Biases for emergent communication in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13111–13121.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2149–2159.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STaR: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling (COLM)*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations (ICLR)*.

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.

Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. 2019. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 693–701.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T. Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, and 22 others. 2024. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*.

Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. In *International Conference on Learning Representations (ICLR)*, pages 10840–10862.

Xiangyuan Xue, Yifan Zhou, Guibin Zhang, Zaibin Zhang, Yijiang Li, Chen Zhang, Zhenfei Yin, Philip Torr, Wanli Ouyang, and Lei Bai. 2026. CoMAS: Co-evolving multi-agent systems via interaction rewards. In *International Conference on Learning Representations (ICLR)*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15476–15488.

A Algorithm and Decision Rule

A.1 Debate + SFT Loop Pseudocode

This appendix gives the exact round-by-round Debate+SFT procedure used in the main experiments.

Algorithm 1 Debate + SFT self-distillation loop

Require: base θ_0; task dist. $\mathcal{D}$; R=5; N=3; personas $\{p_i\}$; filter $\mathcal{F}$

$\phi_0 \leftarrow \varnothing$ ▷ no adapter at round 0
for $r = 0, \ldots, R-1$ **do**
  Sample $x_{1:500} \sim \mathcal{D}$ with seed $s + r \cdot 1009$
  for each x **do**
    $m_i^{(1)} \leftarrow \text{Gen}(x, p_i; \theta_0 + \phi_r)$ for i=1..N
    $m_i^{(2)} \leftarrow \text{Critique}(x, p_i, m_{-i}^{(1)}; \theta_0 + \phi_r)$
    $a_i \leftarrow \text{Parse}(m_i^{(2)})$; $\hat{a} \leftarrow \text{Majority}(\{a_i\})$
  end for
  $\mathcal{T}_r \leftarrow \{(x, m^{(2)}) : \mathcal{F}(\{a_i\}, \text{gold}(x))\}$
  $\phi_{r+1} \leftarrow \text{LoRA-SFT}(\theta_0, \mathcal{T}_r)$
end for

A.2 Pre-Registered Classification Rule

This appendix states the original pre-registered decision rule used to audit trajectory-level communication regimes. The main text uses reader-facing labels (weak refinement, harmful homogenization, and mild inflation) for the discovery trajectories; these labels summarize both the rule outputs and the accompanying effect-size interpretation, rather than replacing the slope-level criteria. The five terms in the table are the only formal rule outputs. The three main-text labels are descriptive summaries that combine the applicable rule output with accuracy direction, message signals, and the matched-Solo contrast; they are not additional rule categories or one-to-one aliases.

Classification	**Conditions** (arm × round slope)
Compression	$\beta_{\text{len}} < 0$, $\beta_{\text{acc}} \geq -0.02$, $\beta_{\text{MATTR}} \geq -0.03$
Collapse	$\beta_{\text{len}} < 0$, $\beta_{\text{acc}} < 0$
Inflation	$\beta_{\text{len}} > 0$ (at $p < .05$, $d \geq 0.5$)
Temp-Sensitive	$\lvert\beta_{\text{MATTR}}\rvert \geq 0.03$ with β_{acc} opposite sign
Inconclusive	None of the above at corrected $p < .05$.

Table 5: Pre-registered decision rule. Each (task, arm) cell is classified after BH-FDR correction over the 18-test grid.

B Robustness Controls and Ablations

This appendix groups the auxiliary controls that test whether the main claim could be explained by train-size effects, persona wording, or temperature schedule alone.

B.1 Volume and Persona Controls

Condition	**R4 Acc**	**R4 MATTR**	n_{train}
Debate (ablation baseline)	.378 (.330–.427)	.690	113
Debate (identical-p)	.543 (.436–.650)	.653	241
Solo (ablation baseline, n=500)	.492 (.443–.541)	.618	242
Solo (matched, n=115)	.510 (.411–.609)	.591	62

Table 6: GSM-Hard robustness controls on an ablation subset; rows are separate from the 5-seed aggregate in Table 1.

Table 6 rules out two simple explanations for the GSM-Hard decline in a controlled ablation subset. Volume-matched Solo remains above the corresponding Debate baseline, so the failure is not caused only by fewer accepted examples after the unanimous-correct training filter. Identical-persona Debate improves accuracy while lowering MATTR, showing that persona differentiation can amplify harmful communication but is not the sole source of the main effect.

B.2 Temperature-Swap Control

Schedule	**ΔAcc**	**ΔMATTR**	**ΔS-BLEU**	n_{train}
0.7/0.3 (default)	−.057	+.053	−.097	113
0.3/0.7 (swapped)	+.035	+.049	−.112	148

Table 7: Temperature-swap ablation on GSM-Hard Debate (R0→R4 change) in the ablation subset.

Table 7 shows that inverting the temperature schedule can reverse the direction of GSM-Hard accuracy while preserving a communication-quality shift. This is why we describe GSM-Hard as temperature-sensitive rather than as a universal debate collapse.

B.3 Per-Agent Role Diagnostics

Figure 3 shows that agent roles can produce distinct lexical-diversity profiles within Debate. Under the identical-persona ablation, per-agent spread narrows, indicating that persona roles shape communication style even when they do not fully determine the endpoint accuracy pattern.

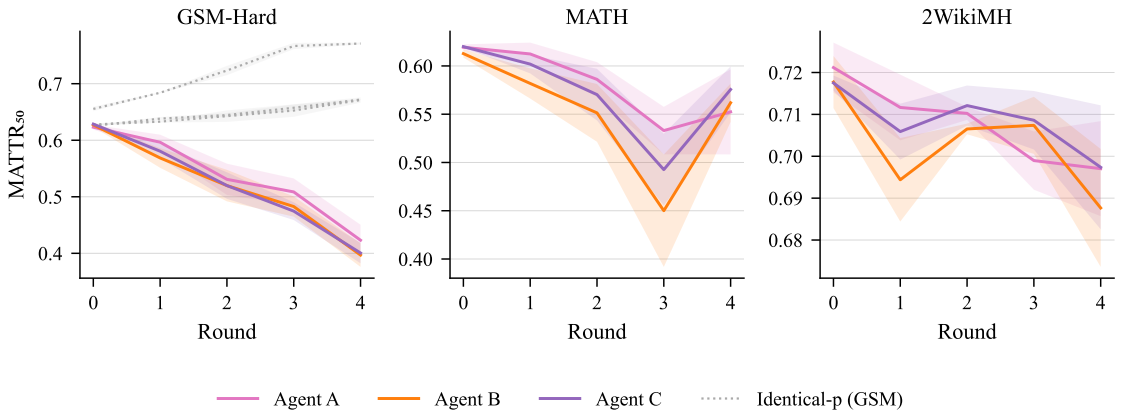


Figure 3: Per-agent MATTR trajectories under differentiated and identical personas.

B.4 Diagnosis-Guided Fallback Sanity Check

As a lightweight utility check, we evaluate a post-hoc fallback policy using only existing round-level metrics. The policy observes Debate through round 2 and falls back to Solo at round 4 when at least three of four warning signals fire: length increase, Self-BLEU increase, MATTR decrease, and unanimous rate decrease. This analysis does not retrain models and does not use round-4 accuracy to choose the arm.

Task	Debate	Solo	Fallback	vs. Debate	vs. Solo
MATH	.167	.229	.188	+.021	−.041
GSM-Hard	.242	.365	.365	+.123	.000
2WikiMHQA	.070	.124	.114	+.044	−.010
Macro	.160	.239	.222	+.063	−.017

Table 8: Post-hoc diagnosis-guided fallback using early communication signals.

Table 8 is a utility sanity check, not a new training algorithm. The diagnostic signal can flag when Debate should be avoided, improving macro accuracy over always-Debate by 6.3pp, but the current task suite still favors always-Solo overall. A fully online fallback policy would require rerunning later rounds after the switch and is left to future work.

C External Validation and Supporting Details

C.1 Cross-Model Evidence: Llama-3.1-8B

To test whether the diagnostic patterns are specific to Qwen2.5-7B, we rerun the discovery-task grid on Llama-3.1-8B-Instruct using the same Debate+SFT pipeline (fresh LoRA, 5 rounds, seeds 17 and 42). The table reports the cross-model cells used in the appendix analysis; MATH Debate is shown for seed 17, while the other rows use seeds 17 and 42.

Task	Arm	Seeds	R0 Acc	R4 Acc	ΔAcc
GSM-Hard	Debate	17,42	.250	.195	−.055
GSM-Hard	Solo	17,42	.255	.348	+.093
MATH	Debate	17	.110	.095	−.015
MATH	Solo	17,42	.335	.325	−.010
2WikiMHQA	Debate	17,42	.158	.193	+.035
2WikiMHQA	Solo	17,42	.160	.303	+.143

Table 9: Llama-3.1-8B-Instruct cross-model accuracy. MATH Debate reports seed 17; the other rows report seeds 17 and 42.

Task	Seeds	R4 Len	ΔLen	R4 MATTR	R4 S-BLEU
GSM-Hard Debate	17,42	302.0	−14.1	.733	.356
MATH Debate	17	305.3	+1.1	.679	.522
2WikiMHQA Debate	17,42	321.9	+74.8	.742	.546

Table 10: Round-4 Debate communication statistics for Llama-3.1-8B.

Tables 9 and 10 show that the Llama results preserve the main comparative warning: Debate does not dominate matched Solo, even when peer messages are available. GSM-Hard reproduces the clearest dissociation, with Debate decreasing by 5.5pp while Solo improves by 9.3pp. 2WikiMHQA is not a harmful-collapse replication, but it still shows a large Solo advantage at round 4: Debate improves by 3.5pp, whereas Solo improves by 14.3pp. Its Debate channel also inflates sharply, with length rising by 74.8 tokens and Self-BLEU increasing to .546. We therefore treat the Llama results as directional support rather than as a separate main result.

A seed-level check further suggests that the GSM-Hard drop is associated with reduced clean-majority coverage, matching the Qwen trend that

communication can reduce the amount of usable consensus even when individual arithmetic remains recoverable.

C.2 Validation Tasks on Related Benchmarks

This appendix summarizes three pilot validation tasks used to test whether the main regimes recur on related benchmarks beyond the discovery set. These runs are supporting checks rather than a second discovery set. The task sources are `openai_humaneval`, `wics/strategy-qa`, and `tau/commonsense_qa`.

Task	Pilot signal	Regime
HumanEval	length -25%, acc. +10pp	Compression
StrategyQA	acc. -4.3pp	Temp-sensitive
CommonsenseQA (CSQA)	length +22%, S-BLEU .42→.58	Inflation

Table 11: Pilot validation tasks on related benchmarks.

Discovery tasks use 5 seeds and support full trajectory analysis. The task-aware T-CIC layer is reported separately for all three discovery tasks on round-4 adapters. The related tasks use 2 seeds and are used only to check whether similar patterns appear beyond the main task set.

Tier	Seeds	Role
Discovery	5	main table and trajectory estimates
Validation	2	family-level recurrence on related tasks

Table 12: Validation-task provenance and evidence-tier roles.

C.3 Hyperparameter Configuration

Shared unless noted: main model Qwen2.5-7B-Instruct; cross-model model Llama-3.1-8B-Instruct; LoRA r=16, α=32, `q_proj/v_proj`, no bias, dropout 0.0. Training uses AdamW, learning rate 2×10^{-4}, cosine schedule, 2 epochs, batch size 2, and bf16. Inference uses vLLM v1, temperatures 0.7/0.3, and top-p 0.95. Max new tokens are 384 for GSM-Hard/2WikiMHQA and 320 for Qwen MATH; max sequence lengths are 4096 for GSM-Hard/2WikiMHQA and 3072 for MATH. Debate uses 3 agents, 2 turns, 500 train and 200 eval samples per round, and 5 rounds. Main seeds are {17, 42, 2024, 1337, 7}; ablations use 2–3 seeds. Hardware is NVIDIA A800 80 GB; the temperature-swap ablation uses 0.3/0.7.

D Qualitative Model Outputs

This final appendix provides compact, reader-facing examples of the model outputs behind the diagnostic labels. We intentionally summarize the raw transcripts rather than printing full generations: several raw outputs contain hundreds of repeated answer strings or prompt leakage, which is itself a useful failure signal but obscures the comparison when reproduced verbatim.

MATH: weak refinement	
Input	Find the number of positive divisors of 2002.
Turn-1 messages	Agent A and Agent C immediately identify the prime factorization $2002 = 2 \cdot 7 \cdot 11 \cdot 13$ and answer 16. Agent B gives a longer step-by-step derivation but truncates before the final expression.
Turn-2 output	After observing peer messages, at least two agents preserve the correct answer and the same divisor-count rule, but the round still contains prompt leakage and copied fragments. This is answer-level recovery rather than clean message-level refinement.
Diagnosis	Weak refinement. The peer channel preserves the majority-correct answer despite residual prompt leakage. The example is useful precisely because the answer signal improves more than the message morphology does.

Table 13: Compact MATH output example. The full raw transcript is omitted because the diagnostic point is the stabilized final reasoning, not the repeated prompt fragments.

GSM-Hard: harmful homogenization	
Input	Jamal's phone holds 1,646,864 photographs and can hold 6 times as many photographs as Brittany's phone. Brittany's capacity is 50 times the number of birds in Jamal's duck photograph. How many ducks are in the photograph?
Ref. calc.	Brittany's capacity is $1{,}646{,}864/6 \approx 274{,}477.33$; the implied bird count is $274{,}477.33/50 \approx 5{,}489.55$. The sample is arithmetically fragile; the diagnostic issue is whether debate amplifies a shortcut or degeneration.
Turn-1 messages	Agent B introduces an unsupported shortcut that maps the problem to `300`. Agent C instead derives a value near 5,489 but then repeats the answer many times. The message channel already contains both arithmetic confusion and degeneracy.
Turn-2 output	Agent A collapses into a repeated string: `Answer: 300` dozens of times. Agent B separately repeats a different answer, while Agent C copies earlier flawed reasoning. The debate increases repetition and disagreement rather than yielding a stable answer.
Diagnosis	Harmful repetition. This local case illustrates degenerative copying and within-agent repetition, the qualitative behavior behind the aggregate harmful-homogenization signal. Endpoint accuracy alone hides this failure mode until the transcript is inspected.

Table 14: Compact GSM-Hard output example. This replaces the raw appendix log that previously filled pages with repeated `Answer: 300` lines.

2WikiMHQA: mild inflation	
Input	Are Adirondack Regional Airport and Marcos Juarez Airport located in the same country?
Turn-1 messages	Agent A and Agent B answer *No*, identifying the airports as being in the United States and Argentina. Agent C contradicts itself, stating that both are in the United States while also placing Marcos Juarez in Argentina, then answers *Yes*.
Turn-2 output	The outputs contain the correct final answer (*No*), but one critique contains a contradictory geography sentence and the contamination persists into turn 2 as agents quote or paraphrase peers and include instruction fragments.
Diagnosis	Mild inflation. The answer can remain correct, yet the message channel becomes verbose and partly contaminated by copied or contradictory text. This supports the paper's claim that message morphology must be monitored separately from endpoint accuracy.

Table 15: Compact 2WikiMHQA output example showing answer recovery together with message inflation.