PEERCOT: STRUCTURED MULTI-AGENT CHAIN-OF-THOUGHT COLLABORATION FOR ERROR LOCALIZATION IN LLM REASONING

Isha Chaturvedi
Independent Researcher
chaturvedi.isha6@gmail.com

Rhys Llewellyn-Jones*
The Pingry School
rhys.m.llewellynjones@gmail.com

Sage Rain Schaffer*
Independent Researcher
schaffersage@gmail.com

ABSTRACT

Large Language Model (LLM) agents exhibit emergent reasoning abilities through debate, critique, and self-reflection. However, most multi-agent systems only exchange final outputs between agents, which limits transparency and hinders the ability to diagnose and improve reasoning processes. PeerCoT introduces a structured and symmetric Chain-of-Thought (CoT) exchange protocol. In this system, peer agents transparently share reasoning traces, provide labeled critiques, and perform minimal-edit revisions before aggregation. This structure enables explicit measurement of process-level error-type identification, called Error Localization Success (ELS), within an agent's reasoning.

We introduce and release *AQUA-RAT-Corrupted*[1] and *GSM8K-Corrupted*[2], structured benchmarks synthetically designed to evaluate error localization and correction in multi-agent reasoning. PeerCoT achieves 64.1% accuracy in AQUA-RAT-Corrupted and 53.15% in GSM8K-Corrupted. PeerCoT maintains competitive accuracy and transparency compared to the baseline models while providing explicit error taxonomy, critique, and ELS. Beyond outcome-level performance, the structured critique protocol corrects 30.43% of initially incorrect solutions in merged outputs. By aligning cooperative critique with fine-grained reasoning supervision, PeerCoT introduces explicit error identification in collaborative reasoning.

1 INTRODUCTION

Modern Large Language Model (LLM) agents can compartmentalize tasks, verify work, and reach consensus. However, most multi-agent settings obscure the reasoning behind their results. Agents typically exchange final outputs rather than intermediate logic, limiting interpretability and error tracing. When agents collaborate without the exposure of their chain of thought (CoT) steps and reasoning, downstream reviewers, whether human or machine, see the reasoning process as a black box. This creates practical consequences: (i) weak support for step-level supervision and repairs; (ii) difficulty assessing peer calibration on sub-skills; and (iii) failure modes where agreement is mistaken for correctness. Recent studies show that multi-agent debates can decrease accuracy in heterogeneous groups by compelling models towards erroneous answers via social-conformity dynamics (Wynn et al., 2025). This evidence promotes protocols that prioritize lucid, step-level exchanges over persuasion dynamics. Thus, rather than primarily targeting higher accuracy, PeerCoT emphasizes error diagnosability in collaborative reasoning.

*These authors contributed equally to this work

[1] https://huggingface.co/datasets/RhysLJ/aqua_rat_corrupted

[2] https://huggingface.co/datasets/RhysLJ/gsm8k_corrupted.

PeerCoT is a collaborative reasoning protocol designed to make intermediate reasoning the primary object of exchange between agents. Two peers independently produce CoTs, exchange labeled critiques, and revise their reasoning before aggregation by a merging model. The protocol is symmetric (no fixed critic/solver roles), transparent (full CoT visibility), and lightweight (prompt-only, no fine-tuning), which collectively enhance the flexibility and interpretability of the reasoning process. PeerCoT instantiates three personas: an Expert Agent (low temperature, 0.3) who emphasizes precision and step-by-step logic; a Curious Student (high temperature, 0.9) who explores hypothesis diversity; and a Merger Agent (low temperature, 0.3) who consolidates revised traces. Revised traces are consolidated here, balancing rigor and exploration to enable transparency while countering single-agent bias.

PeerCoT critiques are step-level, with minimal edits, and labeled by error type. To maintain stability and interpretability, critiques are limited to minimal edits. The merging model preserves full CoT visibility, allowing reviewers to trace how evidence flows through critique, revision, and synthesis. PeerCoT targets reasoning tasks where intermediate steps matter (arithmetic, symbolic, and logic problems), and complements debate by providing a structured framework for error diagnosis and correction. Its core objective is to surface and repair concrete reasoning defects, not persuade a judge. No new pretraining or Reinforcement Learning (RL) objectives are proposed.

The primary contributions of this work are:

- **Structured Collaborative Reasoning:** Two agents exchange critiques over CoT traces, revise their reasoning, and merge results through a third aggregator model.
- **Synthetic Evaluation Benchmark:** We introduce and release structured evaluation benchmarks, AQUA-RAT-Corrupted and GSM8K-Corrupted synthetically constructed from AQUA-RAT (Ling et al., 2017) and GSM8K (Cobbe et al., 2021), respectively, with controlled error injection across four reasoning failure types: arithmetic, omission, logical inconsistency, and ambiguity.
- **Performance:** PeerCoT achieves 64.1% accuracy on AQUA-RAT-Corrupted and 53.1% on GSM8K-Corrupted, maintaining competitive performance relative to single-agent baselines and ablation variants.
- **Interpretability and Diagnostics:** PeerCoT introduces Error Localization Success (ELS), a process-level metric that measures how effectively peer critiques identify faulty reasoning steps, enabling process-level analysis.
- **Constrained Critique Design**: PeerCoT employs a single-round, bidirectional critique–revision mechanism with minimal-edit constraints, providing a controlled setting for studying cooperative reasoning dynamics.

2 Related Work

Collaborative Multi-Agent Systems. Classical MARL explores coordination and shared policy learning (Lowe et al. (2017); Foerster et al. (2018)), typically operating in latent state spaces rather than natural-language reasoning traces. These approaches often employ centralized critics with decentralized actors to estimate contribution and optimize joint reward. In contrast, PeerCoT focuses on interpretable, language-level collaboration, emphasizing shared reasoning visibility rather than reward-based credit assignment.

Debate and Verification. Adversarial debate enhances factuality and reasoning through structured multi-round argument exchange (Du et al., 2023). Debate frameworks often include intermediate justifications, but reasoning traces are generally free-form rather than standardized, making it difficult to trace or audit step-level logic. Debate performance can degrade in heterogeneous agent groups due to social conformity effects (Wynn et al., 2025). Critic–solver and self-refinement frameworks (Shinn et al., 2023; Madaan et al., 2023) impose asymmetric roles in which one model evaluates another's output. Process supervision trains models using step-level feedback rather than only outcome labels, improving reliability on multi-step math reasoning (Lightman et al., 2023). PeerCoT introduces a symmetric, bidirectional reasoning protocol where peers exchange and critique explicit Chain-of-Thought (CoT) steps with labeled error types and perform collaborative minimal-edit revisions. Structured CoT sharing improves interpretability and reduces blind-spot propagation.

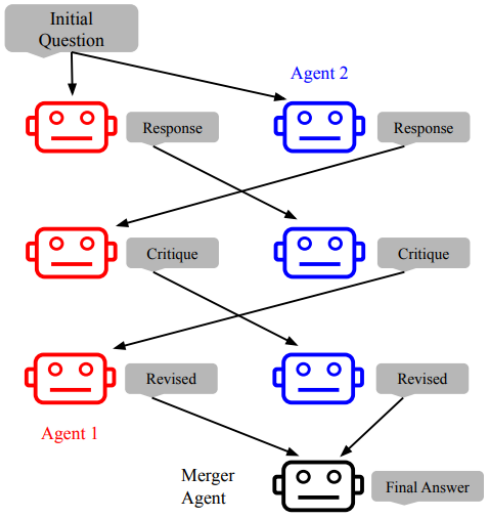

Figure 1: Two agents independently generate CoTs, exchange critiques, revise, and aggregate outputs.

Reasoning Techniques. Chain-of-Thought prompting (Wei et al. (2022)), Tree-of-Thought reasoning (Yao et al. (2023)), and reflective reasoning methods (Hsing (2025)) improve single-agent reasoning through structured multi-step thinking or self-correction. Complementary to these, *self-consistency* samples multiple CoT paths and aggregates their consensus to boost final accuracy (Wang et al., 2022), while *program-aided* approaches offload solution steps to an external interpreter, using LLMs primarily for decomposition (Gao et al., 2022). PeerCoT extends this space to multi-agent settings with shared and revisable reasoning traces that enhance interpretability and support explicit error localization.

Structured Multi-Agent Reasoning. Recent frameworks such as Cochain (Zhao et al. (2025)), Layered CoT (Sanwal (2025)), and Collaborative Language Agents (Zhang et al. (2025)) advance multi-agent reasoning via knowledge sharing or hierarchical goal planning. PeerCoT differs by emphasizing interactive critique and iterative revision between peers, fostering mutual calibration rather than static review.

3 METHOD

Implementation. We implement PeerCoT via prompting with GPT-3.5-Turbo. Each agent: (i) independently generates a CoT; (ii) critiques the peer's CoT with four labeled errors (E1-E4) and brief justification, with redundant or conflicting critiques filtered to reduce noise; (iii) revises its CoT using actionable points and revalidates the reasoning chain to ensure coherence; (iv) and normalizes the revised output to participate in aggregation, yielding the final answer A^*, with all the intermediate traces logged. These implementation steps are designed to reduce noise and maintain consistency.

3.1 PROBLEM SETUP

Given a prompt P, agents f_A and f_B, and an aggregation agent f_M, produce a final answer A^* via collaborative reasoning while maintaining separate, inspectable traces. We define agent f_A as an expert agent using system prompts and is defined with a low temperature, 0.3. This is designed to help create clearer and more arithmetic precise responses. We define agent f_B as a curious student using system prompts with a higher temperature, 0.9. This is designed to help create responses covering a wider range of topics and to create longer, more detailed responses. Finally, we define agent f_M as an arbiter with system prompts and a low temperature, 0.3, to help aggregate the responses and return information in an extractable manner.

3.2 Initial Responses

Agents independently generate their responses to the prompt:

$$\mathrm{CoT}_A^{(0)} = f_A(P), \tag{1}$$

$$\mathrm{CoT}_B^{(0)} = f_B(P). \tag{2}$$

3.3 Critique Generation

Agents annotate peer CoTs for the four error types: **E1** (Arithmetic), **E2** (Omission), **E3** (Logical inconsistency), **E4** (Ambiguity).

$$\mathrm{Critique}_A = f_A(P, \mathrm{CoT}_B^{(0)}), \tag{3}$$

$$\mathrm{Critique}_B = f_B(P, \mathrm{CoT}_A^{(0)}). \tag{4}$$

Critiques must point to specific steps and propose a minimal fix, enabling targeted edits rather than full rewrites.

3.4 Error Taxonomy

We define four interpretable error types for peer critique, each linked to a specific reasoning failure:

- **E1: Arithmetic error** - A numerical or symbolic slip in an otherwise valid step (e.g., $7 \times 8 = 54$).
- **E2: Omission** - A missing or skipped logical step, case, or constraint necessary for completeness.
- **E3: Logical inconsistency** - A contradiction or invalid inference (e.g., applying an incorrect rule or reversing a sign).
- **E4: Ambiguity** - An underspecified or unclear step that allows multiple interpretations (e.g., vague references or undefined variables).

Each critique identifies the faulty step(s) and assigns one of these labels to make reasoning errors explicit and measurable. This step-level supervision signal is aligned with recent evidence that process supervision can outperform outcome-only feedback on multi-step math reasoning (Lightman et al., 2023).

3.5 Critique Exchange

Agents then revise their original responses with the other agents' critique, improving their answer. This includes a verification pass to increase accuracy.

$$\mathrm{CoT}_A^{(1)} = f_A(P, \mathrm{CoT}_A^{(0)} \mathrm{Critique}_B), \tag{5}$$

$$\mathrm{CoT}_B^{(1)} = f_B(P, \mathrm{CoT}_B^{(0)} \mathrm{Critique}_A). \tag{6}$$

3.6 Answer Aggregation

A third agent f_M, called the Merger Agent, aggregates both revised CoTs:

$$A^* = f_M(P, \mathrm{CoT}_A^{(1)}, \mathrm{CoT}_B^{(1)}). \tag{7}$$

This creates the final answer, for which we use an extraction function to find a definitive answer. CoTs are normalized to reduce bias in the performance of the Merger Agent.

3.7 Design Choice

Rather than implementing feedback loops for the critique agent to make continuous improvement, the exchange of CoT only occurs once. This design choice is intended to reduce noise and over-correction, fostering cleaner ELS scores, maintaining interpretability, and minimizing computational costs.

3.8 ILLUSTRATIVE EXAMPLE

Questions: What is $12 - 4 + 3$?

Agent A: $12 + 4 = 16$ (**E3**); $16 + 3 = \mathbf{19}$.
Agent B: $12 - 4 = \mathbf{8}$ (**E2**).

Agent A: You made an omission error (**E2**), forgetting the last step of adding 3.
Agent B: You made a logical inconsistency (**E3**) error and changed the minus sign to a plus sign in step 1.

Agent A: $12 - 4 = 8$; $8 + 3 = \mathbf{11}$.
Agent B: $12 - 4 = 8$; $8 + 3 = \mathbf{11}$.

Agent M: $12 - 4 + 3 = \mathbf{11}$.

4 EVALUATION

4.1 DATASETS

We use the GSM8K and AQUA-RAT datasets, which consist of multi-step arithmetic and algebra reasoning problems. These benchmarks are selected for their structured reasoning chains and multi-step aspects, which make them suitable for evaluating both accuracy and intermediate error localization.

Synthetic Dataset. We construct and release two synthetically corrupted reasoning datasets derived from GSM8K and AQUA-RAT. Each rationale is injected with one of the four reasoning error types (Section 3.4). This is to test whether models can identify and repair distinct reasoning failures. This setup allows us to benchmark PEERCOT's ability to localize and correct specific error types and to conduct targeted ablations across them.

Dataset Creation. We generate the synthetic error-injected datasets using GPT-4o (OpenAI, 2024). For each question–rationale pair in GSM8K and AQUA-RAT, an error type (E1–E4) is randomly assigned. A corruption prompt is constructed with explicit instructions for introducing the corresponding error while preserving the original task structure and context (original question and the correct rationale). This process maintains fidelity to the source datasets while introducing systematic reasoning perturbations.

Error Distribution. Table 1 summarizes the distribution of injected errors across datasets. The counts vary slightly across categories because the error types are assigned randomly at the rationale level. The resulting GSM8K-Corrupted and AQUA-RAT-Corrupted together provide controlled benchmarks for assessing reasoning behavior across injected error types, and self-correction.

Table 1: Distribution of injected reasoning errors (E1–E4) across synthetic datasets. Each rationale in GSM8K and AQUA-RAT is randomly assigned one error type.

Dataset	E1	E2	E3	E4	Total
GSM8K	1,897	1,867	1,855	1,854	7,473
AQUA-RAT	5,065	5,049	5,038	4,848	20,000

4.2 METRICS

We use the following four metrics to evaluate a response from PeerCoT and the baselines.

- **Accuracy**
  The percentage of final correct answers within each 2000-question benchmark.

- **Improvement Rate (IR)**
  The percentage of cases where both initial agents f_A and f_B are incorrect, but the merged output A^* is correct.

$$\text{IR} = \frac{\#(f_A \text{and} f_B \text{ wrong, and } A^* \text{ correct})}{\#(f_A \text{and} f_B \text{ wrong})} \times 100$$

- **Error Localization Success (ELS)**
  The percentage of critiques on incorrect CoTs where the peer critique correctly identifies the error.

$$\text{ELS} = \frac{\#(\text{critiques correctly identifying the error})}{\#(\text{critiques on incorrect CoTs})} \times 100$$

- **Transparency Score (TS)**
  GPT-5 OpenAI (2025) rates the clarity and traceability of the model's Chain-of-Thought (CoT) on a 1–5 Likert scale (Likert (1932)). A higher score indicates more explicit, interpretable reasoning with well-structured intermediate steps. The evaluation rubric is defined as follows.

Score	Description
1	Unclear, disorganized, or missing reasoning steps.
2	Some logic present but fragmented or vague.
3	Mostly clear reasoning but skips steps or contains redundancy.
4	Step-by-step reasoning that is traceable and interpretable.
5	Exceptionally clear, concise, and fully transparent reasoning chain.

All baselines and ablations use identical definitions of Agents in the paper.

4.3 BASELINES

To evaluate the performance of PeerCoT, we construct three baselines based on comparable multi-agent reasoning paradigms.

- **Solo CoT**
  A single agent performs reasoning and produces a final answer using standard Chain-of-Thought (CoT) prompting, without feedback or collaboration. This baseline isolates individual reasoning performance, contrasting with PeerCoT's multi-agent critique and revision process.
- **Multi-agent Debate**
  Two agents engage in adversarial dialogue, exchanging final answers-not full CoTs-over multiple rounds to reach consensus. Each refines its response based on the opponent's critique, without access to reasoning steps. Unlike PeerCoT's transparent CoT sharing, this setup emphasizes persuasion over collaborative reasoning.
- **One-way Reflection (e.g., Reflexion)**
  One agent generates a solution while another provides one-way feedback without sharing its own CoT. This asymmetric setup tests the impact of external critique without mutual reasoning or self-reflection, contrasting PeerCoT's symmetric exchange and collaborative revision.

4.4 ABLATION STUDIES

In addition to the baselines, we created four ablation studies. These help measure the contributions of different aspects of the method. We observe how our overall accuracy is affected by removing certain parts and aspects of our method.

- **Ablation 1**: No CoT Sharing (Independent Agents Only)
  Purpose: Establish a baseline where both agents operate independently without observing

or critiquing each other's reasoning. This isolates individual problem-solving ability and removes collaboration effects for comparison against PeerCoT.
Setup: Both agents receive the same prompt and independently produce answers without access to one another's reasoning. No critique or reflection occurs between agents, and the final output is selected arbitrarily between the two answers.

- **Ablation 2**: One-Way CoT Sharing
  Purpose: Examine whether symmetric reasoning exchange offers advantages over asymmetric supervision, and assess whether mutual critique, where both agents generate and review solutions, outperforms a one sided feedback setup.
  Setup: Agent A has access to and critiques Agent B's Chain-of-Thought (CoT), but not vice versa. As a result, Agent B can refine its reasoning based on feedback, while Agent A cannot. This configuration isolates the contribution of bidirectional reflection by contrasting active mutual critique with passive, one-way evaluation.

- **Ablation 3**: No Critique (Just CoT Visibility)
  Purpose: Evaluate whether passive exposure to a peer's CoT enhances reasoning performance, or if active critique and error labeling are necessary for improvement.
  Setup: Both agents can view each other's CoTs but do not provide explicit feedback or error annotations. This configuration isolates the effect of simple CoT visibility, testing whether observation alone, without structured critique, can support reasoning refinement.

- **Ablation 4**: Single-Turn (No History)
  Purpose: Measure the effect of removing feedback and memory while retaining symmetric CoT visibility. This variant isolates the contribution of multi-turn context to sustained reasoning improvement.
  Setup: Single-round CoT exchange for mutual critique, without history or iterative refinement. This model bridges adversarial debate frameworks, which exchange only final claims and critic solver paradigms, which are one-directional, by preserving a cooperative but non-iterative reasoning structure.

5 RESULTS

All methods are implemented using ChatGPT-3.5-Turbo and evaluated across the four metrics applicable to each configuration. Performance summaries for PeerCoT, its baselines, and ablation variants on the AQUA-RAT benchmark are presented in Table 2, while results on the GSM8K benchmark are reported in Table 3. For the purpose of this study, we consider Improvement Rate (IR) and Error Localization Success (ELS) as secondary metrics.

Accuracy. PeerCoT demonstrates competitive accuracy across both the AQUA-RAT-Corrupted and GSM8K-Corrupted benchmarks (Table 2, Table 3). On AQUA-RAT-Corrupted, PeerCoT attains 64.10%, surpassing Solo CoT (62.45%) but slightly below Multi-Agent Debate (64.85%) and One-Way Reflection (65.20%). Among the ablations, PeerCoT exceeds the Single-Turn No History variant (62.70%) but is marginally outperformed by No CoT Sharing (67.10%), One-Way Sharing (65.50%), and No Critique (64.10%). This trend suggests that while full bidirectional critique enhances interpretability, simpler coordination schemes can occasionally yield higher raw accuracy due to reduced interaction noise. On GSM8K-Corrupted, PeerCoT achieves 53.15%, slightly below Solo CoT (57.25%) and Multi-Agent Debate (56.00%) but comparable to One-Way Reflection (51.20%). Overall, PeerCoT maintains competitive accuracy while enabling measurable self-correction and error localization.

Transparency Score. PeerCoT achieves a Transparency Score of 3.54 out of 5 on the AQUA-RAT-Corrupted dataset, lower than Solo CoT (3.79) but higher than all other baselines and ablation variants. On GSM8K-Corrupted, PeerCoT reaches 4.21 out of 5, closely matching Solo CoT (4.24) and exceeding all other models. The higher transparency of Solo CoT arises from its single-step reasoning process, which produces concise outputs without critique or revision. PeerCoT, in contrast, maintains strong transparency despite multi-agent critique and merging, indicating that collaborative feedback improves clarity and coherence without substantially increasing complexity.

Table 2: Merged results of baseline models and ablation studies on the 2000-question benchmark for the *AQUA-RAT-Corrupted*. Accuracy (%) measures task correctness, IR = Improvement Rate, ELS = Error Localization Success, TS = Transparency Score, MS = Misread Percentage. Sharing Mechanism and Visibility describe the interaction setup between agents.

Model / Setting	Accuracy (%)	IR (%)	ELS (%)	TS	MS (%)	Sharing Mechanism	Visibility
Solo CoT	62.45±1.08	–	–	3.79/5.0	6.10±0.54	None	None
Multi-agent debate	64.85±1.07	19.33±0.88	–	2.34/5.0	0.4±0.14	Bidirectional (no critique)	Shared
One-way reflection	65.20±1.07	–	–	2.40/5.0	1.05±0.23	One-way sharing	Partial
A1: No CoT Sharing	67.10±1.05	3.17±0.39	–	2.66/5.0	0.2±0.10	None (agents independent)	None
A2: One-Way Sharing	65.50±1.06	26.15±0.98	–	2.72/5.0	0.65±0.18	Asymmetric (one shares, other observes)	Partial
A3: No Critique (Just CoT Visibility)	64.10±1.07	30.84±1.03	–	2.76/5.0	0.30±0.12	None (no feedback, visibility only)	Shared
A4: Single-Turn (No History)	62.70±1.08	25.54±0.98	9.07±0.64	2.00/5.0	0.45±0.15	One-round exchange, no iteration	Symmetric
PeerCoT	**64.10±1.07**	**30.43±1.03**	**11.18±0.70**	**3.54/5.0**	**0.3±0.12**	**Bidirectional with critique**	**Full**

Table 3: Results across baselines and the full PEERCOT model on a 2000-question benchmark for the *GSM8K-Corrupted*. IR = Improvement Rate. ELS = Error Localization Success. TS = Transparency Score. MS = Misread Percentage

Model	Accuracy(%)	IR(%)	ELS(%)	TS	MS%	Sharing Mechanism	Visibility
Solo CoT	57.25±1.11	–	–	4.24/5.0	9.40±0.65	None	None
Multi-agent debate	56.00±1.11	7.76±0.60	–	4.11/5.0	1.45±0.27	Bidirectional (no critique)	Shared
One-way reflection	51.20±1.12	–	–	3.98/5.0	0.95±0.22	One-way sharing	Partial
A1: No CoT Sharing	52.30±1.12	4.35±0.46	–	3.98/5.0	0.95±0.22	None (agents independent)	None
A2: One-Way Sharing	54.95±1.11	8.76±0.63	–	4.13/5.0	3.80±0.43	Asymmetric (one shares, other observes)	Partial
A3: No Critique (Just CoT Visibility)	57.90±1.10	8.79±0.63	–	4.17/5.0	1.40±0.26	None (no feedback, visibility only)	Shared
A4: Single-Turn (No History)	48.55±1.12	6.30±0.54	6.45±0.55	4.11/5.0	3.65±0.42	One-round exchange, no iteration	Symmetric
PeerCoT	**53.15±1.12**	**10.49±0.69**	**7.41±0.59**	**4.21/5.0**	**1.50±0.27**	**Bidirectional with critique**	**Full**

Secondary metrics. Secondary metrics provide a more detailed view of PeerCoT's reasoning performance beyond overall accuracy. The Improvement Rate (IR) captures how often PeerCoT successfully refines or corrects an initially flawed rationale through multi-agent critique (Section 4.2). PeerCoT achieves IR of 30.43% on AQUA-RAT-Corrupted and 10.49% on GSM8K-Corrupted, indicating frequent correction of initially incorrect solutions. Note that IR cannot be computed for single-answer baselines such as Solo CoT and One-way Reflection, which lack iterative feedback or merging steps. The Error Localization Success (ELS) metric represents the percentage of incorrect CoTs where agents correctly identify the faulty step. ELS scores across both datasets provide quantitative evidence of step-level error identification during collaborative reasoning. The Misread Percentage (MS%) reflects cases where the merging agent's response cannot be parsed by the result-extraction function—typically due to formatting issues or invalid output structure. Although unrelated to reasoning ability, all misread samples are conservatively treated as incorrect during evaluation to maintain consistency across metrics.

6 DISCUSSION AND CONCLUSION

Discussion. While PeerCoT does not always exceed baseline models in raw accuracy, this outcome reflects a design trade-off rather than a limitation. The PeerCoT framework naturally integrates both corrective and disruptive peer interventions, where agents can make wrong answers right and right answers wrong. This is PeerCoT's central strength: it enables us to observe the process and identify how and why errors arise. PeerCoT transforms LLM logic chains into traceable and interpretable reasoning steps, revealing why models succeed or fail. For instance, Table 4 shows that PeerCoT's merger agent introduces more corruptions than improvements, illustrating the inherent tension between aggregation and reasoning precision.

Conclusion. This paper introduces PeerCoT, a structured collaborative reasoning protocol for large language models (LLMs) in which two agents independently generate solutions, exchange and critique each other's reasoning, revise their responses, and then aggregate them through a third merger agent. We evaluate PeerCoT on synthetically corrupted reasoning datasets that we release publicly, constructed from GSM8K and AQUA-RAT and augmented with controlled error injection to model four different reasoning failure types. This setup enables systematic assessment of reasoning behavior and interpretability across collaborative configurations. Across both datasets (Tables 2 and 3),

Table 4: Stage-wise counts of corruptions and improvements across protocol variants. Entries show per-stage tallies within each pipeline step.

Protocol	Stage Corruption Counts	Stage Improve Counts
Base 1: Solo CoT	--	merger: 1299
Base 2: Multi-Agent Debate	revision_A: 83, revision_B: 90, merger: 101	revision_A: 20, revision_B: 5, merger: 13
Base 3: One-Way Reflection	revision_A: 110, merger: 1	revision_A: 112
A1: No CoT Sharing	merger: 149	merger: 3
A2: One-Way Sharing	revision_B: 48, merger: 222	revision_B: 37, merger: 10
A3: No Critique	revision_A: 46, revision_B: 71, merger: 137	revision_A: 29, revision_B: 30, merger: 3
A4: Single-Turn	revision_A: 71, revision_B: 37, merger: 214	revision_A: 11, revision_B: 14, merger: 9
PeerCoT	revision_A: 79, revision_B: 31, merger: 203	revision_A: 33, revision_B: 33, merger: 12

PeerCoT achieves competitive accuracy while maintaining strong and competitive interpretability. On AQUA-RAT-Corrupted, PeerCoT attains 64.1% accuracy—higher than Solo CoT (62.5%) but slightly below Multi-Agent Debate and One-Way Reflection. This demonstrates that structured bidirectional critique can achieve comparable performance to debate frameworks, without unconstrained multi-turn exchanges. On *GSM8K-Corrupted*, PeerCoT achieves 53.1% accuracy and maintains high transparency (4.21/5). Notably, PeerCoT introduces and reports Error Localization Success (11.18% on AQUA-RAT-Corrupted, 7.41% on GSM8K-Corrupted), highlighting its diagnostic capacity to identify reasoning errors directly.

Table 4 further illustrates this interpretability–accuracy balance. Although the merger stage introduces notable corruption, the critique–revision exchanges remain constructive, showing that structured peer feedback enhances reasoning reliability even when aggregation introduces noise. Moreover, PeerCoT's architecture naturally extends to multi-round critique–revision cycles, allowing iterative refinement when additional rounds yield measurable benefit, without requiring continuous debate.

Overall, PeerCoT demonstrates that bidirectional critique and revision yield an interpretable form of multi-agent reasoning. By combining transparency, diagnostic granularity, and structured collaboration, PeerCoT offers a scalable framework for analyzing and improving reasoning processes in large language models.

7 Limitations and Future Work

While PeerCoT demonstrates strong performance and interpretability benefits, several limitations remain. First, the multi-agent structure inherently increases computational cost. Each task requires multiple model invocations, distinct prompting stages, and aggregation, resulting in longer processing times compared to single-agent reasoning. Second, the transparency scores are evaluated using a GPT-5 as a proxy for human judgment. Although this provides a consistent and scalable metric, it cannot fully capture human nuances in assessing reasoning clarity. Achieving the most accurate interpretability assessments would require human evaluation of all responses, which is currently impractical at scale.

In future work, we plan to expand evaluation using a larger portion of our synthetic error datasets, allowing a more comprehensive analysis of PeerCoT's robustness across reasoning error types. We also aim to test the framework across different LLMs to examine model sensitivity and generalization. Further extensions may explore adaptive critique cycles and selective CoT sharing to reduce computational overhead while maintaining interpretability and accuracy.

References

Karl Cobbe, Vineet Kosaraju, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Yilun Du, Shuang Li, et al. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022. URL https://arxiv.org/abs/2211.10435.

Nicole Hsing. Mirror: Modular internal processing for personalized safety in llm dialogue. *arXiv preprint arXiv:2506.00430*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

R. Likert. A technique for the measurement of attitudes. *Archives of Psychology, 22(140), 1–55*, 1932.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Aman Madaan, Niket Tandon, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

OpenAI. Gpt-4o. https://openai.com/, 2024. Large language model accessed via OpenAI API.

OpenAI. Gpt-5 technical overview. https://openai.com/research/gpt-5, 2025. Model used for Chain-of-Thought clarity evaluation in this study.

Manish Sanwal. Layered chain-of-thought prompting for multi-agent llm systems: A comprehensive approach to explainable large language models. *arXiv preprint arXiv:2501.18645*, 2025. URL https://arxiv.org/abs/2501.18645.

Noah Shinn, Francesco Cassano, et al. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025. URL https://arxiv.org/abs/2509.05396.

Shunyu Yao, Dian Yu, Jeffrey Zhao, et al. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Cong Zhang, Xin Deik Goh, Dexun Li, Hao Zhang, and Yong Liu. Planning with multi-constraints via collaborative language agents. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025. URL https://aclanthology.org/2025.coling-main.672/.

Jiaxing Zhao, Hongbin Xie, Yuzhen Lei, Xuan Song, Zhuoran Shi, Lianxin Li, et al. Cochain: Balancing insufficient and excessive collaboration in llm agent workflows. *arXiv preprint arXiv:2505.10936*, 2025. URL https://arxiv.org/abs/2505.10936.