EPISTEMIC CONTEXT LEARNING: BUILDING TRUST THE RIGHT WAY IN LLM MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent systems provide individual agents with peer-generated information, yet no principled framework exists for how a single agent should optimally utilize this information. Existing approaches either design better communication protocols—a system-level concern orthogonal to individual decision-making—or aggregate peer responses without regard to source reliability. We formalize a fundamentally different, *individual-level* problem: *history-aware reference*, where an agent conditions its reasoning on interaction history to estimate peer reliability and selectively learn from trustworthy peers. Diagnostic analysis reveals that current LLMs suffer from two critical failure modes under this formulation: inability to extract trust signals from history and blind conformity to peer consensus. To address these, we develop *Epistemic Context Learning* (ECL), a two-stage framework that decouples trust estimation from final reasoning, optimized with auxiliary reinforcement learning rewards for explicit peer reliability modeling. Experiments on MMLU-Pro and GPQA show that ECL consistently and significantly outperforms history-agnostic baselines across 8 models from 4 families. Analysis confirms that trust modeling accuracy is strongly correlated with final answer quality, and that ECL generalizes robustly across diverse multi-agent configurations.

1 INTRODUCTION

Large Language Models (LLMs) have achieved significant success in complex reasoning and agentic automation (Wei et al., 2022; Yao et al., 2023), facilitating the deployment of collaborative multi-agent systems across various real-world applications (Hong et al., 2024; Qian et al., 2024). When an agent in a multi-agent (MA) system receives responses from peers, it faces a fundamental decision: how should it utilize this information to improve its own reasoning? This is not a question about how agents should be organized or what protocol they should follow—those are system-level design concerns studied extensively in prior work (Wu et al., 2024b; Hong et al., 2024). Instead, it is an *individual-level* optimization problem: given a fixed set of peers whose outputs are already available, what is the best strategy for a single agent to selectively incorporate peer information?

This question is critically under-studied. Existing multi-agent methods predominantly focus on designing better communication protocols, debate mechanisms, or agent topologies (Park et al., 2025; Chen et al., 2024; Zhang et al., 2025b). They assume that improving the *structure* of interaction is the path to better collective outcomes. However, even under an ideal protocol, an individual agent still must decide how much to trust each peer's response—and current agents are remarkably poor at this. LLMs are prone to blind conformity to majority opinions (Zhu et al., 2025; Cho et al., 2025) and are easily misled by plausible but incorrect reasoning (Shi et al., 2023; Wu et al., 2024a). Recent work has further identified multiple failure modes of multi-agent debate itself, including error propagation, sycophancy, and degradation over rounds (Chen et al., 2025; Zhang et al., 2025c; Wynn et al., 2025). These failures are not protocol failures—they are failures of individual information utilization.

At the core of these failures is a mismatch between task difficulty and the agent's available signals. When its own knowledge is insufficient, an agent is implicitly required to judge peer correctness from the *content* of a single response. Under such conditions, a confidently stated hallucination easily dominates a terse correct answer, because surface-level plausibility is the only discriminative signal (Lightman et al., 2024; Zhao et al., 2025). We argue that this is the fundamental limitation of any *content-only* evaluation: it asks the agent to verify claims it cannot independently assess.

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

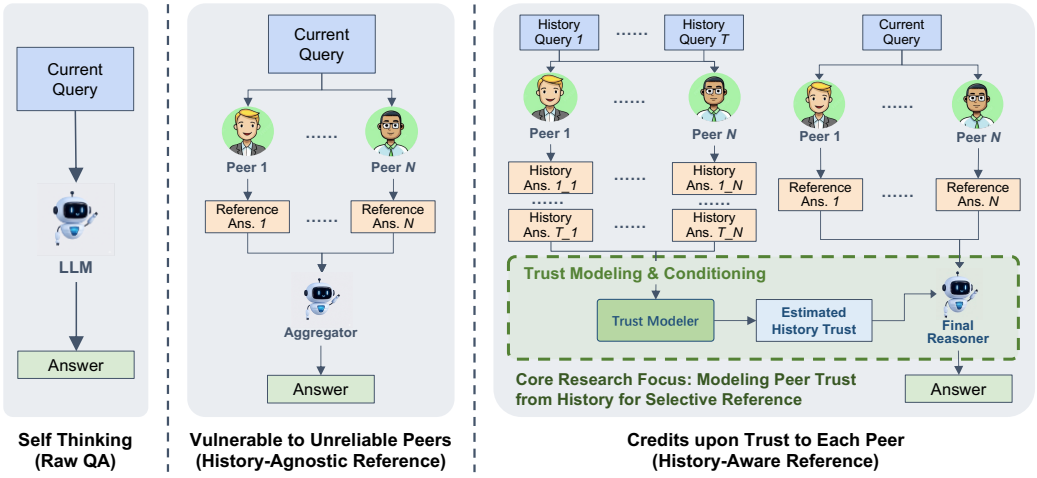

Figure 1: How an individual agent processes peer information. (Left) The agent reasons independently without peer input. (Middle) The agent consults peers but treats all responses equally, remaining vulnerable to plausible misinformation. (Right) The agent conditions its reference on estimated peer reliability from interaction history, enabling selective trust.

When content evaluation is unreliable, the problem naturally shifts from evaluating *what is said* to evaluating *who is saying it*. Historical reliability provides an epistemic prior that is far easier to estimate than the validity of a specific reasoning trace. We formalize this insight as *history-aware reference* (Section 3): the agent conditions its reasoning on interaction history to derive trust estimates, then uses these estimates to selectively weight peer information. As illustrated in Figure 1, this transforms the agent from a passive aggregator into an active, trust-conditioned reasoner.

We emphasize that our contribution is orthogonal to multi-agent protocol design. ECL does not prescribe how agents should communicate; it prescribes how an individual agent should *use* whatever information the protocol makes available. It can be applied on top of any existing MA protocol, as it only modifies how the receiving agent processes incoming peer responses.

Diagnostic analysis (Section 4) reveals that current LLMs, even after RL training, suffer from two failure modes under history-aware reference: (1) **Lack of Historical Trust**—they still ignore history and rely on superficial cues in the current round; and (2) **Blind Conformity**—they over-rely on peer consensus at the expense of internal reasoning. Driven by these findings, we develop *Epistemic Context Learning (ECL)*, a two-stage framework that architecturally decouples trust estimation from final reasoning. The first stage processes *only* the history to produce a peer reliability profile; the second stage conditions on this profile alongside current peer responses to generate the final answer. We further introduce *Peer Recognition Reward (PRR)*, an auxiliary RL signal that provides direct supervision for trust estimation in Stage 1.

Our main contributions are:

- We formalize *history-aware reference*, an individual-level problem for optimal peer information utilization, and identify two critical failure modes of current LLMs under this formulation.
- We propose ECL, a structured reasoning framework with auxiliary RL supervision that decouples trust estimation from final reasoning, preventing shortcut learning and reward hacking.
- Experiments across 8 models from 4 families show ECL consistently and significantly outperforms history-agnostic baselines. Analysis confirms trust modeling accuracy is strongly correlated with final reasoning quality and that ECL generalizes across diverse MA configurations.

2 Related Work

LLM-based multi-agent systems. LLM-based MAS have been applied across diverse scenarios (Wu et al., 2024b; Hong et al., 2024; Qian et al., 2024). Multi-agent debate can improve reasoning

through iterative argumentation (Du et al., 2024; Liang et al., 2024), but recent work has identified fundamental failure modes: debate often collapses into simple ensembling (Chen et al., 2025; Ai et al., 2025), suffers from error propagation and conformity (Wynn et al., 2025; Song et al., 2026), and can degrade performance over rounds (Zhang et al., 2025c). Rather than proposing a better protocol to mitigate these issues, we address their root cause: individual agents lack a principled mechanism for evaluating peer reliability.

System-level optimization vs. individual-level utilization. Several works optimize MAS at the system level—deciding who communicates with whom. SelfOrg (Chen et al., 2024) constructs dynamic topologies; G-Designer (Zhang et al., 2025b) uses GNNs for communication graphs; MAS-GPT (Ye et al., 2025) trains LLMs to build MAS configurations. These operate at a different level of abstraction: they design the system, while we optimize the individual agent's strategy *within* the system. ECL is complementary and can be applied on top of any topology or protocol.

Conformity, sycophancy, and truth discovery. LLMs exhibit sycophancy (Sharma et al., 2024), conformity (Zhu et al., 2025; Cho et al., 2025), and sensitivity to social pressure (Kim et al., 2024; Weng et al., 2025; Choi et al., 2025). Existing fixes operate within the current round. We resolve these vulnerabilities by grounding evaluation in historical trust rather than current-round plausibility. Our trust estimation is related to truth discovery (Li et al., 2016), but ECL goes beyond label aggregation: estimated trust conditions the agent's full reasoning process, enabling selective adoption of peer reasoning patterns.

3 Problem: History-Aware Reference

We consider an LLM-based MAS involving N agents $\mathcal{A} \equiv \{A_1, \ldots, A_N\}$, and focus on a specific **current agent** $A_{\text{curr}} \in \mathcal{A}$. All other agents $\mathcal{P} \equiv \mathcal{A} \setminus \{A_{\text{curr}}\}$ are **peer agents**.

The task of *history-aware reference* is defined on a dataset $\mathcal{D} \equiv \{(Q_j, Y_j, \mathcal{H}_j)\}_{j=1}^{M}$, where:

- Q_j is the current query;
- Y_j is the ground-truth answer; and
- $\mathcal{H}_j \equiv \{(Q_j^{(k)}, \mathcal{R}_j^{(k)})\}_{k=1}^{T_j}$ is a sequence of T_j historical rounds of peer responses.

At the current round, peers provide responses $\mathcal{R}_j \equiv \{\hat{y}_{j,p} = A_p(Q_j) \mid A_p \in \mathcal{P}\}$, and the current agent produces $\hat{Y}_j = A_{\text{curr}}(\mathcal{H}_j, Q_j, \mathcal{R}_j)$ conditioned on the entire context. Performance is evaluated by comparing $\hat{Y}_j$ with Y_j.

Standard multi-agent aggregation operates in a history-agnostic manner: $\hat{Y}_j = A_{\text{curr}}(Q_j, \mathcal{R}_j)$. This fails when peer reliability is non-uniform and cannot be discerned from the current interaction alone (e.g., confident hallucinations). History-aware reference introduces $\mathcal{H}_j$ as a critical conditioning variable: the agent must extract latent beliefs regarding peer reliability from $\mathcal{H}_j$ and use these beliefs to modulate its utilization of $\mathcal{R}_j$.

4 Diagnostic Analysis

Before presenting our method, we diagnose how current LLMs handle history-aware reference. We construct a controlled environment with two peers ($|\mathcal{P}| = 2$): one consistently reliable, one consistently unreliable, using 5 rounds of history. We train Qwen 2.5-3B and Llama 3.2-3B with GRPO (Shao et al., 2024) under three information configurations to isolate the impact of peer context:

- **Single Agent (SA):** The agent receives only the current query without any access to peers—equivalent to standard RL training.
- **MA-Outcome:** The agent receives both the interaction history and peer responses in the current round, but only their final answers without reasoning steps.
- **MA-Reasoning:** The agent additionally receives peer reasoning traces in the current round, allowing it to evaluate the quality of peer rationales.

Table 1: Diagnostic probing: Flipping Identity (Flip) and All Wrong (All-W) settings reveal that RL-trained models ignore history and exhibit blind conformity.

Datasets	Models	Metric	MA-Outcome	MA-Reasoning
Math500	Qwen 2.5-3B	Acc.	73.0%	86.5%
		Acc. (Flip)	62.2% (-10.8%)	81.1% (-5.4%)
		Acc. (All-W)	5.4% (-67.6%)	48.7% (-37.8%)
	Llama 3.2-3B	Acc.	43.2%	86.5%
		Acc. (Flip)	51.4% (+8.2%)	83.8% (-2.7%)
		Acc. (All-W)	13.5% (-29.7%)	32.4% (-54.1%)
LiveCode	Qwen 2.5-3B	Acc.	75.7%	86.5%
		Acc. (Flip)	67.6% (-8.1%)	89.2% (+2.7%)
		Acc. (All-W)	12.2% (-63.5%)	32.4% (-54.1%)
	Llama 3.2-3B	Acc.	59.5%	83.8%
		Acc. (Flip)	51.4% (-8.1%)	85.1% (+1.3%)
		Acc. (All-W)	5.4% (-54.1%)	20.3% (-63.5%)

RL training with peer context generally improves performance over SA (full results in Appendix A.1), but it remains unclear whether this improvement stems from genuine trust estimated via historical track records or merely from aggregating plausible current-round answers. We introduce two diagnostic settings to disentangle these factors.

Finding 1: Lack of Historical Trust. In the **Flipping Identity (Flip)** setting, we invert peer reliability at test time: the historically reliable peer provides incorrect answers while the unreliable peer provides correct ones. If models truly condition on history, they would trust the now-incorrect peer and accuracy should drop sharply. Conversely, a model that evaluates peers solely by current-round content would be unaffected, since the swap does not alter the quality of each peer's reasoning in the current round. As shown in Table 1, accuracy remains largely stable under Flip (e.g., Qwen 2.5-3B drops only 5.4% in MA-Reasoning on Math500, and Llama 3.2-3B even *improves*). This indicates that models *ignore history* and rely on current-round cues—they learn to aggregate plausible answers, not to track reliability.

Finding 2: Blind Conformity. In the **All Wrong (All-W)** setting, both peers answer incorrectly at test time. If models maintain epistemic autonomy, they should fall back on internal knowledge. Instead, Table 1 shows that accuracy collapses dramatically—below 15% for MA-Outcome and below 50% for MA-Reasoning—confirming that standard RL produces agents that over-rely on peer consensus at the expense of internal reasoning. Notably, these models achieve 37–65% standalone accuracy under the SA condition (Appendix A.1), demonstrating sufficient independent capability; the collapse to single-digit accuracy in MA-Outcome reveals that unanimously wrong peers override the model's own correct reasoning entirely.

These two failure modes motivate our approach along two directions: (1) *Structured architectures* that force explicit history processing, inspired by human reputation establishment (Behrens et al., 2008; Sperber et al., 2010); and (2) *Auxiliary supervision* that provides denser signals for trust estimation beyond sparse outcome rewards.

5 Epistemic Context Learning

We propose *Epistemic Context Learning (ECL)*, a two-stage framework that architecturally decouples trust estimation from final reasoning, optimized with auxiliary RL rewards.

Two-Stage Structured Reasoning. We decompose the reasoning process into two stages:

- **Stage 1: Trust Estimation.** The model receives *only* the history $\mathcal{H}_j$ and extracts an epistemic belief profile $B_j = A_i(\mathcal{H}_j)$. The current query Q_j and peer responses $\mathcal{R}_j$ are *withheld*, creating an **information bottleneck** (Tishby et al., 2000) that forces the model to compress historical evidence into a trust prior rather than exploiting current-round shortcuts.
- **Stage 2: Trust-Conditioned Reasoning.** The raw history is replaced by the extracted belief profile. The model receives B_j, the current query Q_j, and peer responses $\mathcal{R}_j$, then generates the final answer $\hat{Y}_j = A_i(B_j, Q_j, \mathcal{R}_j)$, conditioning its aggregation on the trust prior.

Table 3: Controlled comparison across architectures and rewards (Qwen 2.5-3B, $|\mathcal{P}| = 2$). ECL (E) achieves the best performance by combining structured reasoning with direct trust supervision.

Datasets	MA-Outcome			MA-Reasoning		
	1S (RL)	ECL (I)	ECL (E)	1S (RL)	ECL (I)	ECL (E)
Math500	73.0%	75.7%	81.1%	86.5%	94.6%	100.0%
LiveCode	75.7%	77.0%	89.2%	86.5%	91.9%	100.0%

Even without training, the two-stage (2S) pipeline significantly outperforms single-stage (1S) processing on base LLMs, as shown in Table 2. This confirms that structurally separating history analysis from current-round reasoning introduces a beneficial inductive bias.

Table 2: Base LLMs (no training) on 1S vs. 2S pipeline ($|\mathcal{P}| = 2$). Decoupling yields large gains.

Datasets	Models	MA-Outcome		MA-Reasoning	
		1S	2S	1S	2S
Math500	Qwen 2.5-3B	67.6%	78.4%	56.7%	67.6%
	Llama 3.2-3B	43.2%	59.4%	48.7%	59.4%
LiveCode	Qwen 2.5-3B	50.0%	75.7%	66.2%	74.3%
	Llama 3.2-3B	37.8%	63.5%	64.9%	71.6%

RL with Auxiliary Supervision. The 2-stage structure provides architectural inductive bias, but how to optimize both stages effectively remains open. Directly propagating the final outcome reward (OR) to Stage 1 can cause instability, so we introduce an auxiliary *Peer Recognition Reward* (PRR) for Stage 1. We prompt the model to conclude Stage 1 with `The most reliable peer is: <PEER_NAME>` and compare against the ground truth:

$$r_{\text{PRR}} = \begin{cases} 1.0 & \text{If } \langle \text{PEER_NAME} \rangle = \text{Reliable Peer} \\ 0.0 & \text{Otherwise} \end{cases}$$

Stage 1 is trained with PRR and Stage 2 with OR, so each stage receives supervision depending only on its own output. We denote the PRR-supervised variant as **ECL (E)** (**E**xplicit trust supervision), the OR-only variant as **ECL (I)** (**I**mplicit), and the single-stage approach as **1S**.

Table 3 compares the three variants. ECL (I) modestly outperforms 1S (e.g., 77.0% vs. 75.7% on LiveCode), confirming the architectural benefit. ECL (E) achieves substantially larger gains (5–10% on average), reaching 100% on both datasets under MA-Reasoning. This validates that direct supervision for trust estimation in Stage 1 is critical.

Why PRR requires two-stage decoupling. Naïvely applying PRR to single-stage RL causes **reward hacking**: the model exploits current-round cues to predict the reliable peer without genuinely reasoning from history, leading to performance *degradation* (e.g., 73.0%→62.2% on Math500; see Appendix A.2). Our two-stage design blocks this shortcut by conditioning Stage 1 solely on history.

6 Evaluation

6.1 Settings

LLMs. We test 8 models from 4 families:

- **Qwen 3 (Yang et al., 2025):** 4B, 8B, and 30B.[1]
- **Deepseek V3.2 (DeepSeek-AI et al., 2025)**.
- **GPT-5-mini, GPT-5.2 (OpenAI, 2025)**.
- **Gemini 3 Flash & Pro (Google Gemini Team, 2025)**.

Table 4: Dataset statistics.

Dataset	History	Train	Test
MMLU-Pro	100	810	90
GPQA	50	358	40

We finetune Qwen 3-4B and Qwen 3-8B (instruct versions) with RL, and evaluate larger models at inference time without RL training. All results are averaged over 5 random seeds; we report means in the main table and include standard deviations in Appendix A.4. To further reduce variance on the smaller GPQA test set, we additionally evaluate three representative models (DeepSeek V3.2, GPT-5-mini, Gemini 3 Flash) on an enlarged 90-sample test split; results are in Appendix A.5.

[1] Specific versions: Qwen3-4B-Instruct-2507, Qwen3-8B, Qwen3-30B-Thinking-2507.

Datasets. We use MMLU-Pro (Wang et al., 2024) (commonsense reasoning) and GPQA (Rein et al., 2024) (scientific reasoning). For MMLU-Pro, we randomly sample 1000 from the original 12000 questions; for GPQA, we use all 448 questions. We set aside 100 (MMLU-Pro) and 50 (GPQA) questions as the history pool, and split the rest 9:1 into training and test sets. Statistics are in Table 4.

MA context. We use $|\mathcal{P}| = 4$ peers and $|\mathcal{H}_j| = 5$ history rounds, and create two settings:

- **Natural:** 4 different LLMs as peers (Gemini 3 Flash as the strong one; Llama 3.1-8B, Gemma 3-4B, Qwen 3-0.6B as peers), all generating their responses via pure sampling. This simulates varying but non-adversarial peer reliability.
- **Adversarial:** All 4 peers built with Qwen3-30B-A3B, one providing correct answers with reasoning, the other three injecting subtle errors. This tests robustness against deceptive agents.

Research questions and baselines. Existing aggregation methods focus on current-round responses. We organize our comparisons and analyses around four research questions:

- **RQ 1: Does historical trust improve peer information use?** We compare ECL with Single Agent (SA; independent reasoning) and Aggregation (AG; current-round peers only) in Table 5.
- **RQ 2: How does ECL establish and use trust?**
  - **RQ 2.1:** Can agents identify reliable peers? We measure PRR at test time and during RL (Table 6, Figure 2).
  - **RQ 2.2:** Do trust estimates guide decisions? We flip peer reliability and compare accuracy under correct/incorrect identification (Appendix A.6, Table 7).
  - **RQ 2.3:** Why two stages? Tables 2 and 3 already establish Stage 1's benefit over history-aware 1S; Copy Reliable Peer (CRP), which copies the peer selected in Stage 1, tests Stage 2's contribution (Table 8).
- **RQ 3: How does ECL use inference compute?**
  - **RQ 3.1:** Can extra reasoning replace historical trust? We compare with 1–3 rounds of single-agent reflection (Table 9).
  - **RQ 3.2:** What inference overhead does ECL incur? We measure wall-clock time against AG (Appendix A.7).
- **RQ 4: How broadly does ECL apply?**
  - **RQ 4.1:** Does it work without RL? We compare base-model ECL with AG (Appendix A.3).
  - **RQ 4.2:** Does it generalize across peer counts and history lengths? We vary both and compare with AG (Tables 10 and 11).

6.2 Main Results

RQ 1: Historical trust improves peer information use. Table 5 shows that both ECL variants outperform history-agnostic AG in most settings, with ECL (E) achieving the best result in the majority of comparisons. Historical reliability enables agents to identify and prioritize useful peer information. Five-seed evaluations show significant improvements over AG ($p < 0.05$, paired t-test; Appendix A.4), and the pattern persists on the enlarged 90-sample GPQA test set (Appendix A.5).

Historical trust strengthens resistance to deceptive peers. Current-round aggregation can underperform independent reasoning: on adversarial GPQA, Qwen 3-30B drops from 72.5% (SA) to 70.0% (AG), while ECL (E) reaches 80.0% in MA-Outcome. Trust also helps frontier models distinguish convincing errors from reliable information: Gemini 3 Flash improves from 84.4% to 97.8% on adversarial MMLU-Pro. These gains show the value of historical evidence when current-round content is misleading.

Benefits span model scales and training regimes. RL-tuned Qwen 3-4B/8B use peer context to bridge knowledge gaps, while larger models benefit from ECL at inference time alone across all four model families. Gains depend on the information peers contribute: Gemini 3 Flash and Pro benefit less in the natural setting, where the strongest peer is itself Gemini 3 Flash, but improve substantially with adversarial peers; Gemini 3 Pro reaches 100% on adversarial GPQA with ECL (E).

Table 5: Main results on MMLU-Pro and GPQA. ECL (E) requires explicit peer recognition in Stage 1 and rewards it via PRR during RL (for Qwen 3-4B/8B); ECL (I) uses implicit trust estimation. For Qwen 3-4B/8B, results are from RL-tuned models; others are inference-only. Best result per row across MA variants is **bolded**. Results are averaged over 5 seeds (std. in Appendix A.4).

Datasets	Models	SA		MA-Outcome			MA-Reasoning		
		Base	RL	AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)
Task Setting: Natural									
MMLU-Pro	Qwen 3-4B	–	70.0%	67.8%	**72.2%**	71.1%	78.9%	**85.6%**	84.4%
	Qwen 3-8B	–	71.1%	66.7%	65.6%	**77.8%**	71.1%	72.2%	**76.7%**
	Qwen 3-30B	75.6%	–	73.3%	**77.8%**	**77.8%**	82.2%	82.2%	**86.7%**
	Deepseek V3.2	80.0%	–	81.1%	**86.7%**	84.4%	83.3%	83.3%	83.3%
	GPT-5-mini	80.0%	–	82.2%	84.4%	**85.6%**	**86.7%**	85.6%	**86.7%**
	GPT-5.2	84.4%	–	85.6%	83.3%	**86.7%**	85.6%	84.4%	**86.7%**
	Gemini 3 Flash	85.6%	–	82.2%	**86.7%**	85.6%	83.3%	84.4%	**85.6%**
	Gemini 3 Pro	87.8%	–	**87.8%**	85.6%	85.6%	85.6%	85.6%	**86.7%**
GPQA	Qwen 3-4B	–	55.0%	42.5%	62.5%	**70.0%**	**82.5%**	**82.5%**	77.5%
	Qwen 3-8B	–	52.5%	57.5%	55.0%	**65.0%**	**77.5%**	75.0%	**77.5%**
	Qwen 3-30B	72.5%	–	65.0%	70.0%	**75.0%**	82.5%	**90.0%**	85.0%
	DeepSeek V3.2	77.5%	–	77.5%	75.0%	**85.0%**	85.0%	82.5%	**87.5%**
	GPT-5-mini	82.5%	–	85.0%	80.0%	**92.5%**	90.0%	87.5%	**92.5%**
	GPT-5.2	87.5%	–	90.0%	90.0%	**95.0%**	92.5%	**95.0%**	92.5%
	Gemini 3 Flash	90.0%	–	**92.5%**	90.0%	85.0%	**92.5%**	90.0%	85.0%
	Gemini 3 Pro	90.0%	–	90.0%	**92.5%**	85.0%	92.5%	**95.0%**	**95.0%**
Task Setting: Adversarial									
MMLU-Pro	Qwen 3-4B	–	70.0%	70.0%	**85.6%**	82.2%	78.9%	**90.0%**	**90.0%**
	Qwen 3-8B	–	71.1%	71.1%	77.8%	**83.3%**	71.1%	76.7%	**82.2%**
	Qwen 3-30B	75.6%	–	75.6%	**86.7%**	83.3%	81.1%	85.6%	**92.2%**
	DeepSeek V3.2	80.0%	–	83.3%	**92.2%**	87.8%	80.0%	**93.3%**	91.1%
	GPT-5-mini	80.0%	–	84.4%	85.6%	**87.8%**	84.4%	85.6%	**90.0%**
	GPT-5.2	84.4%	–	85.6%	**88.9%**	**88.9%**	82.2%	**87.8%**	**87.8%**
	Gemini 3 Flash	85.6%	–	84.4%	96.7%	**97.8%**	87.8%	**98.9%**	**98.9%**
	Gemini 3 Pro	87.8%	–	91.1%	**97.8%**	96.7%	87.8%	97.8%	**98.9%**
GPQA	Qwen 3-4B	–	55.0%	47.5%	60.0%	**67.5%**	62.5%	60.0%	**70.0%**
	Qwen 3-8B	–	52.5%	45.0%	**50.0%**	**50.0%**	50.0%	**57.5%**	52.5%
	Qwen 3-30B	72.5%	–	70.0%	72.5%	**80.0%**	65.0%	75.0%	**77.5%**
	DeepSeek V3.2	77.5%	–	77.5%	80.0%	**82.5%**	80.0%	82.5%	**85.0%**
	GPT-5-mini	82.5%	–	**87.5%**	**87.5%**	85.0%	75.0%	82.5%	**85.0%**
	GPT-5.2	87.5%	–	92.5%	90.0%	**97.5%**	87.5%	82.5%	**90.0%**
	Gemini 3 Flash	90.0%	–	87.5%	95.0%	**97.5%**	87.5%	92.5%	**95.0%**
	Gemini 3 Pro	90.0%	–	90.0%	97.5%	**100.0%**	90.0%	**95.0%**	**95.0%**

ECL (I) vs. ECL (E). Explicit identification is particularly effective when peer reliability is polarized: ECL (E) is best or tied-best in 14 of 16 adversarial GPQA model×context comparisons. ECL (I) can better capture graded reliability in the natural setting: Gemini 3 Flash reaches 90.0% vs. 85.0% on GPQA MA-Reasoning, and Gemini 3 Pro reaches 92.5% vs. 85.0% on MA-Outcome. This suggests ECL (E) as the default for sharply differentiated peers, with ECL (I)'s free-form trust profiles offering flexibility for less polarized reliability. Qualitative GPQA cases with RL-trained Qwen 3-4B show how these representations affect reasoning: ECL (I) retains per-peer assessments and can override its highest-rated peer, whereas ECL (E) can carry an erroneous peer selection into the final answer. Appendix A.8.2 traces both stages for two such cases.

6.3 Analysis

Unless otherwise stated, analyses use adversarial peers and ECL (E); full results are in Appendix A.

RQ 2.1: Frontier models identify reliable peers; smaller models learn with RL. We analyze LLMs' ability to model epistemic trust via the PRR values of ECL (E) at test time. As shown in Table 6, on MMLU-Pro all models achieve near-perfect recognition rates. On the harder GPQA dataset, frontier models (DeepSeek V3.2, GPT-5-mini, Gemini 3 Flash) maintain high precision, while smaller models (Qwen 3-4B/8B) score 40-52%, substantially above the 25% random baseline for $|\mathcal{P}| = 4$. Figure 2 shows that PRR values improve steadily during RL training, confirming that trust modeling is a robust capability in large LLMs and a learnable skill for smaller ones.

Table 6: Average r_{PRR} of different models on MMLU-Pro and GPQA.

Datasets	Models	MA-Outcome	MA-Reasoning
MMLU-Pro	Qwen 3-4B	96.7%	98.9%
	Qwen 3-8B	86.7%	82.2%
	DeepSeek V3.2	93.3%	93.3%
	GPT-5-mini	98.9%	98.9%
	Gemini 3 Flash	97.8%	100.0%
GPQA	Qwen 3-4B	52.5%	50.0%
	Qwen 3-8B	40.0%	42.0%
	DeepSeek V3.2	85.0%	90.0%
	GPT-5-mini	90.0%	92.5%
	Gemini 3 Flash	92.5%	92.5%

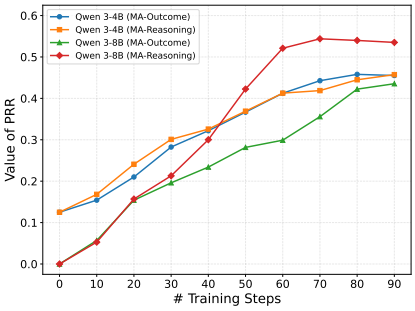

Figure 2: Learning curves of PRR from Qwen 3 on GPQA.

RQ 2.2: Reasoning responds to estimated trust. In the Flip setting (where historically reliable peers become unreliable in the current round at test time), ECL-equipped models show sharp accuracy drops—e.g., Gemini 3 Flash with ECL (I) plummets from 96.7% to 46.7% on MMLU-Pro and from 95.0% to 50.0% on GPQA (Appendix A.6). This degradation confirms models actively condition on the trust prior rather than ignoring it. Furthermore, as shown in Table 7, correct peer identification ($r_{\mathrm{PRR}} = 1$) yields significantly higher accuracy—e.g., 100% vs. 66.7% for Gemini 3 Flash—yet accuracy rarely collapses when trust is misplaced, indicating partial autonomy even when following the wrong peer.

Table 7: Final accuracy conditioned on correct ($r_{\mathrm{PRR}} = 1$) vs. incorrect ($r_{\mathrm{PRR}} = 0$) peer identification on GPQA (adversarial).

Models	MA-Outcome		MA-Reasoning	
	$r_{\mathrm{PRR}} = 1$	$r_{\mathrm{PRR}} = 0$	$r_{\mathrm{PRR}} = 1$	$r_{\mathrm{PRR}} = 0$
Qwen 3-4B	66.7%	57.9%	85.0%	55.0%
Qwen 3-8B	75.0%	33.3%	70.6%	39.1%
DeepSeek V3.2	85.3%	66.7%	86.1%	75.0%
GPT-5-mini	86.1%	75.0%	83.8%	100.0%
Gemini 3 Flash	100.0%	66.7%	100.0%	66.7%

RQ 2.3: Both stages contribute to ECL's gains. Tables 2 and 3 establish the benefit of separate trust estimation over history-aware 1S processing. To examine Stage 2, Table 8 reports Stage 1 identification accuracy (**Peer ID**), *Copy Reliable Peer (CRP)*, and ECL (E). CRP retains the same history-only Stage 1 but directly copies the selected peer's answer. Despite high identification rates (82.5–92.5%), copying remains vulnerable to the selected peer's errors. ECL (E) improves over CRP by 2.5–12.5 percentage points, including 75.0%→87.5% for DeepSeek V3.2, demonstrating the value of reasoning with the trust estimate after peer selection.

Table 8: CRP vs. ECL (E) on GPQA (natural, MA-Reasoning).

Models	Peer ID	CRP	ECL (E)
DeepSeek V3.2	82.5%	75.0%	**87.5%**
Gemini 3 Flash	92.5%	82.5%	**85.0%**
GPT-5.2	87.5%	85.0%	**92.5%**

RQ 3.1: Historical trust beats self-refinement. Against single-agent reflection with $R \in \{1, 2, 3\}$ rounds, ECL (E) exceeds the best reflection accuracy in five of six comparisons and ties in the sixth (Table 9). For example, Gemini 3 Flash reaches at most 86.7% with reflection on MMLU-Pro, versus 98.9% with ECL (E). Historical trust supplies useful evidence about whom to rely on that additional self-refinement does not recover.

Table 9: Single-agent reflection (R rounds) vs. ECL (E) on adversarial MA-Reasoning.

Models	$R = 1$	$R = 2$	$R = 3$	ECL (E)
GPQA				
DeepSeek V3.2	67.5%	67.5%	75.0%	**85.0%**
Gemini 3 Flash	87.5%	92.5%	87.5%	**95.0%**
GPT-5-mini	77.5%	85.0%	85.0%	**85.0%**
MMLU-Pro				
DeepSeek V3.2	75.6%	78.9%	77.8%	**91.1%**
Gemini 3 Flash	85.6%	84.4%	86.7%	**98.9%**
GPT-5-mini	80.0%	80.0%	78.9%	**90.0%**

RQ 3.2: ECL takes 1.4× AG's inference time. ECL's two-stage design increases wall-clock inference time by only 1.4× on average relative to the single-stage AG baseline (Appendix A.7). This overhead is modest compared to the accuracy gains (often 10+ points absolute) and substantially cheaper than multi-round reflection, which at $R = 3$ incurs ∼3× the compute yet yields smaller improvements.

RQ 4.1: ECL benefits base models without RL. ECL also benefits models without RL training. As detailed in Appendix A.3, Qwen 3-4B/8B base models under the ECL pipeline already achieve consistent improvements over AG by leveraging their inherent reasoning ability to summarize peer reliability from history. However, RL becomes essential on harder tasks (e.g., adversarial GPQA),

Table 10: ECL generalizes to different numbers of peers on MMLU-Pro (adversarial, MA-Reasoning).

Models	$\|P\| = 2$			$\|P\| = 3$			$\|P\| = 4$		
	AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)
Qwen 3-4B	82.2%	88.9%	**95.6%**	76.7%	84.4%	**87.8%**	78.9%	**90.0%**	**90.0%**
DeepSeek V3.2	82.2%	88.9%	**94.4%**	80.0%	**94.4%**	**94.4%**	80.0%	**93.3%**	91.1%
GPT-5-mini	84.4%	85.6%	**88.9%**	82.2%	**88.9%**	85.6%	84.4%	85.6%	**90.0%**

Table 11: ECL generalizes to different history lengths on MMLU-Pro (adversarial, MA-Reasoning).

Models	$\|\mathcal{H}_j\| = 2$			$\|\mathcal{H}_j\| = 5$			$\|\mathcal{H}_j\| = 8$		
	AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)
Qwen 3-4B	78.9%	84.4%	**87.8%**	78.9%	**90.0%**	**90.0%**	78.9%	85.6%	**88.9%**
DeepSeek V3.2	83.3%	**88.9%**	86.7%	80.0%	**93.3%**	91.1%	82.2%	**96.7%**	91.1%
GPT-5-mini	80.0%	84.4%	**86.7%**	84.4%	85.6%	**90.0%**	83.3%	**86.7%**	**86.7%**

where base models sometimes fail to recognize reliable peers and ECL can underperform AG, which underscores RL's role in grounding trust modeling under adversarial conditions.

RQ 4.2: ECL generalizes across peer counts and history lengths. We validate that ECL generalizes across MA configurations by varying the number of peers and history rounds on MMLU-Pro (adversarial, MA-Reasoning). As shown in Tables 10 and 11, ECL consistently outperforms AG across $|\mathcal{P}| \in \{2, 3, 4\}$ peers and $|\mathcal{H}_j| \in \{2, 5, 8\}$ history rounds. ECL remains highly effective even with minimal history: most models can still achieve strong performance with only 2 rounds. This demonstrates that minimal interaction evidence suffices for useful trust estimation, and the framework scales robustly regardless of crowd size.

7 Conclusion and Limitations

We formalize *history-aware reference*, an individual-level problem for how a single agent should optimally utilize peer information given access to interaction history—a question orthogonal to multi-agent protocol design. Diagnostic analysis reveals that current LLMs fail to extract trust from history and exhibit blind conformity to peer consensus. To address these failures, we develop *Epistemic Context Learning* (ECL), a two-stage framework that decouples trust estimation from final reasoning, optimized with auxiliary rewards for explicit peer reliability modeling. Experiments across 8 models from 4 families on MMLU-Pro and GPQA demonstrate consistent and significant improvements over history-agnostic baselines, with analysis confirming that trust modeling accuracy strongly correlates with reasoning quality. ECL generalizes robustly across diverse multi-agent configurations, establishing trust-conditioned information utilization as a practical and broadly applicable approach for individual agents in multi-agent systems.

ECL is designed for the regime where peer interaction history is available and peer reliability is approximately stationary—assumptions shared by trust and reputation systems across many domains. Our evaluation adopts controlled data synthesis to enable reproducible ablation over peer reliability configurations, a standard methodology for studying multi-agent dynamics in isolation. These choices are deliberately minimal: our generalization analysis shows that as few as 2 history rounds suffice, and the natural setting confirms effectiveness when no peer is perfectly reliable. Future work includes adaptive trust updating for non-stationary peer behavior, history augmentation strategies for low-resource scenarios, and integration with complementary multi-agent paradigms such as communication topology optimization.

AI Use Statement

We used generative AI tools to assist with implementing selected algorithmic modules, brainstorming and refining the narrative and organization of the Introduction and Related Work sections, and editing the manuscript for clarity and fluency. As part of the experimental setup, we also used LLMs to generate peer responses and reasoning traces, as described in Appendix F. The authors take responsibility for the correctness, originality, and integrity of the final manuscript and implementation, including all AI-assisted content.

Ethics Statement

This work advances the robustness of LLM-based multi-agent systems by enabling agents to assess peer trustworthiness from historical interactions. By guiding the selective use of peer information, ECL helps mitigate sycophancy, conformity, and susceptibility to deceptive or hallucinatory responses, supporting more reliable collaboration in open environments. An ethical concern is that historically reliable peers may later fail, become compromised, or exploit accumulated trust to manipulate other agents. Responsible deployment therefore calls for monitoring changes in peer reliability, updating trust estimates, and preserving agents' capacity for independent reasoning. Developing adaptive trust mechanisms that resist such manipulation is an important direction for future work.

Reproducibility Statement

- Our anonymized code is available at `https://anonymous.4open.science/r/ECL`.
- Section 6 specifies the evaluated models, dataset splits, and experimental settings.
- We provide prompt templates in Appendix C and details of data construction, history sampling, training hyper-parameters, and hardware in Appendix F.

REFERENCES

Rui Ai, Yuqi Pan, David Simchi-Levi, Milind Tambe, and Haifeng Xu. Beyond majority voting: Llm aggregation by leveraging higher-order information. *arXiv preprint arXiv:2510.01499*, 2025. URL https://arxiv.org/abs/2510.01499.

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6929–6948, 2024. URL https://aclanthology.org/2024.findings-emnlp.407/.

Timothy EJ Behrens, Laurence T Hunt, Mark W Woolrich, and Matthew FS Rushworth. Associative learning of social value. *Nature*, 456(7219):245–249, 2008. URL https://doi.org/10.1038/nature07538.

Jiaqi Chen, Yuxian Jiang, Jiachen Lu, and Li Zhang. S-agents: Self-organizing agents in open-ended environments. *arXiv preprint arXiv:2402.04578*, 2024. URL https://arxiv.org/abs/2402.04578.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Optima: Optimizing effectiveness and efficiency for llm-based multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11534–11557, 2025. URL https://arxiv.org/abs/2410.08115.

Sitao Cheng, Xunjian Yin, Ruiwen Zhou, Yuxuan Li, Xinyi Wang, Liangming Pan, William Yang Wang, and Victor Zhong. From atomic to composite: Reinforcement learning enables generalization in complementary reasoning. *arXiv preprint arXiv:2512.01970*, 2025. URL https://arxiv.org/abs/2512.01970.

Young-Min Cho, Sharath Chandra Guntuku, and Lyle Ungar. Herd behavior: Investigating peer influence in llm-based multi-agent systems. *arXiv preprint arXiv:2505.21588*, 2025. URL https://arxiv.org/abs/2505.21588.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. *arXiv preprint arXiv:2510.07517*, 2025. URL https://arxiv.org/abs/2510.07517.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu,

Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. URL https://arxiv.org/abs/2512.02556.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=zj7YuTE4t8.

Google Gemini Team. Gemini 3. 2025. URL https://deepmind.google/models/gemini/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Kyusik Kim, Hyeonseok Jeon, Jeongwoo Ryu, and Bongwon Suh. Will LLMs sink or swim? exploring decision-making under pressure. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11425–11450, 2024. URL https://aclanthology.org/2024.findings-emnlp.668/.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016. URL https://dl.acm.org/doi/10.1145/2897350.2897352.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 17889–17904, 2024. URL https://aclanthology.org/2024.emnlp-main.992/.

Junwei Liao, Muning Wen, Jun Wang, and Weinan Zhang. Marft: Multi-agent reinforcement fine-tuning. *arXiv preprint arXiv:2504.16129*, 2025. URL https://arxiv.org/abs/2504.16129.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022. URL https://aclanthology.org/2022.deelio-1.10/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. URL https://arxiv.org/abs/2503.20783.

OpenAI. Introducing gpt 5. *OpenAI Blogs*, 2025. URL https://openai.com/index/introducing-gpt-5.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS)*, pp. 27730–27744, 2022. URL https://arxiv.org/abs/2203.02155.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E. Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30215–30248, 2025. URL https://aclanthology.org/2025.acl-long.1459/.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, 2024. URL https://aclanthology.org/2024.acl-long.810/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 31210–31227, 2023. URL https://openreview.net/forum?id=JSZmoN03Op.

Maojia Song, Tej Deep Pala, Ruiwen Zhou, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. Measuring and mitigating rapport bias of large language models under multi-agent social interactions. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=gF31wuYdk7.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010. URL https://doi.org/10.1111/j.1468-0017.2010.01394.x.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Proceedings of the 34th Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3008–3021, 2020. URL https://arxiv.org/abs/2009.01325.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023. URL https://aclanthology.org/2023.acl-long.557/.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Proceedings of the 38th Conference on Neural Information Processing Systems 2024 (NeurIPS 2024)*, 2024. URL https://arxiv.org/abs/2406.01574.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=st77ShxP1K.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *The 38th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=WGoCZl2itU.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the 1st Conference on Language Modeling (COLM)*, 2024b. URL https://openreview.net/forum?id=BAakY1hNKS.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2509.05396.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Siheng Chen, and Jing Shao. Mas-gpt: Training llms to build llm-based multi-agent systems. *arXiv preprint arXiv:2503.03686*, 2025. URL https://arxiv.org/abs/2503.03686.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*, 2025a. URL https://arxiv.org/abs/2512.07783.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025b. URL https://arxiv.org/abs/2410.11782.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025c. URL https://arxiv.org/abs/2502.08788.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=xZDW0oejD.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Weinan Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Yutao Qi, et al. Memrl: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*, 2026. URL https://arxiv.org/abs/2601.03192.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan O Arik. Chain of agents: Large language models collaborating on long-context tasks. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024b. URL https://openreview.net/forum?id=LuCLf4BJsr.

Wenting Zhao, Pranjal Aggarwal, Swarnadeep Saha, Asli Celikyilmaz, Jason Weston, and Ilia Kulikov. The majority is not always right: Rl training for solution aggregation. *arXiv preprint arXiv:2509.06870*, 2025. URL https://arxiv.org/abs/2509.06870.

Ruiwen Zhou, Yingxuan Yang, Muning Wen, Ying Wen, Wenhao Wang, Chunling Xi, Guoqiang Xu, Yong Yu, and Weinan Zhang. TRAD: Enhancing llm agents with step-wise thought retrieval and aligned decision. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2024. URL https://dl.acm.org/doi/10.1145/3626772.3657788.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025. URL https://arxiv.org/abs/2506.15841.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, 2025. URL https://aclanthology.org/2025.acl-long.195/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Additional Experimental Results

This section consolidates all supplementary experimental results referenced throughout the main text. Unless otherwise stated, experiments are conducted in adversarial task settings with MA-Reasoning context.

A.1 Diagnostic Analysis: Full Results

Table 12 presents the full diagnostic analysis results referenced in Section 4. The controlled environment uses two peer agents ($|\mathcal{P}| = 2$) with 5 rounds of history on Math500 and LiveCode datasets.

Table 12: Diagnostic analysis: RL consistently strengthens peer response aggregation across both MA context variants.

Datasets	Models	SA		MA-Outcome		MA-Reasoning	
		Base	RL	Base	RL	Base	RL
Math500	Qwen 2.5-3B	51.4%	64.9%	67.6%	73.0%	56.7%	86.5%
	Llama 3.2-3B	51.4%	37.8%	43.2%	43.2%	48.7%	86.5%
LiveCode	Qwen 2.5-3B	66.2%	63.5%	50.0%	75.7%	66.2%	86.5%
	Llama 3.2-3B	23.0%	43.2%	37.8%	59.5%	64.9%	83.8%

A.2 PRR in Single-Stage RL

Table 13 shows that naïvely combining PRR with OR in single-stage RL leads to reward hacking and performance degradation, motivating the two-stage decoupling in ECL.

Table 13: 1S RL with OR + PRR vs. OR only (Qwen 2.5-3B). Adding PRR without decoupling causes reward hacking.

Datasets	MA-Outcome		MA-Reasoning	
	1S (OR)	1S (OR + PRR)	1S (OR)	1S (OR + PRR)
Math500	73.0%	62.2%	86.5%	75.7%
LiveCode	75.7%	55.4%	86.5%	82.4%

A.3 Training-Free ECL on Smaller LLMs

In Section 6 we presented the performances of RL-optimized ECL on small LLMs, i.e. Qwen 3-4B and Qwen 3-8B, and now we show the comparative results between base (RL-free) and RL-optimized models in Table 14.

Our findings are summarized as follows:

- **RL effectiveness and performance saturation.** RL training provides significant performance gains in scenarios where the base model's reasoning or trust-modeling capabilities are underdeveloped. For instance, on GPQA, RL training consistently improves both SA and ECL scores for Qwen 3-8B. However, in settings where the base model already demonstrates high accuracy–such as the reasoning scores on MMLU-Pro–the impact of RL is less pronounced, indicating that the model's potential in those specific configurations is approaching saturation.
- **ECL's utility without training and failure modes.** Even without specialized training, ECL generally outperforms the history-agnostic AG baseline by leveraging the base model's inherent reasoning to summarize peer reliability. Nevertheless, this capability has limits in high-difficulty settings. In the adversarial GPQA task, RL-free models often fail to correctly identify the reliable peer, causing the ECL variant to sometimes underperform the AG baseline (e.g., Qwen 3-8B in MA-Outcome). This underscores the necessity of RL for grounding trust modeling when the distinction between reliable and adversarial peers is subtle.

Table 14: Comparison of Qwen 3-4B and Qwen 3-8B before and after RL training on MMLU-Pro and GPQA datasets. For models without RL training, the difference between ECL (I) and ECL (E) is that the latter restricts the format of stage 1 summary to be the same as PRR required. In each line the best results within two MA context variants are bolded.

Datasets	Models	SA		MA-Outcome			MA-Reasoning		
		Base	RL	AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)
Task Setting: Natural									
MMLU-Pro	Qwen 3-4B (w/o RL)	67.8%	–	66.7%	75.6%	**82.2%**	78.9%	**85.6%**	80.0%
	Qwen 3-4B (w/ RL)	–	70.0%	67.8%	**72.2%**	71.1%	78.9%	**85.6%**	84.4%
	Qwen 3-8B (w/o RL)	65.6%	–	65.6%	64.4%	**66.7%**	67.8%	67.8%	**75.6%**
	Qwen 3-8B (w/ RL)	–	71.1%	66.7%	65.6%	**77.8%**	71.1%	72.2%	**76.7%**
GPQA	Qwen 3-4B (w/o RL)	62.5%	–	55.0%	55.0%	**57.5%**	77.5%	**82.5%**	**82.5%**
	Qwen 3-4B (w/ RL)	–	55.0%	42.5%	62.5%	**70.0%**	82.5%	82.5%	77.5%
	Qwen 3-8B (w/o RL)	47.5%	–	50.0%	55.0%	47.5%	67.5%	80.0%	80.0%
	Qwen 3-8B (w/ RL)	–	52.5%	57.5%	55.0%	**65.0%**	**77.5%**	75.0%	**77.5%**
Task Setting: Adversarial									
MMLU-Pro	Qwen 3-4B (w/o RL)	67.8%	–	73.3%	82.2%	**84.4%**	75.6%	86.7%	**90.0%**
	Qwen 3-4B (w/ RL)	–	70.0%	70.0%	**85.6%**	82.2%	78.9%	**90.0%**	**90.0%**
	Qwen 3-8B (w/o RL)	67.8%	–	60.0%	54.4%	**70.0%**	58.9%	**82.2%**	**82.2%**
	Qwen 3-8B (w/ RL)	–	71.1%	71.1%	77.8%	**83.3%**	71.1%	76.7%	**82.2%**
GPQA	Qwen 3-4B (w/o RL)	60.0%	–	**65.0%**	55.0%	62.5%	**60.0%**	57.5%	57.5%
	Qwen 3-4B (w/ RL)	–	55.0%	47.5%	60.0%	**67.5%**	62.5%	60.0%	**70.0%**
	Qwen 3-8B (w/o RL)	45.0%	–	**42.5%**	37.5%	35.0%	42.5%	45.0%	**47.5%**
	Qwen 3-8B (w/ RL)	–	52.5%	45.0%	**50.0%**	**50.0%**	50.0%	**57.5%**	52.5%

A.4 Full Results with Standard Deviations

We report results averaged over 5 random seeds with standard deviations for representative models on adversarial settings (MA-Reasoning):

Table 15: Results with standard deviations (5 seeds) on adversarial MA-Reasoning settings. ECL advantages are statistically significant ($p < 0.05$, Student's t-test) in bolded cells.

	SA	AG	ECL (I)	ECL (E)
MMLU-Pro				
DeepSeek V3.2	80.0% ± 1.0%	80.0% ± 1.8%	**93.3% ± 1.4%**	**91.1% ± 0.7%**
Gemini 3 Flash	85.6% ± 0.7%	87.8% ± 0.7%	**98.9% ± 1.2%**	**98.9% ± 1.8%**
GPT-5-mini	84.4% ± 1.2%	82.2% ± 1.0%	**87.8% ± 1.4%**	**87.8% ± 0.7%**
GPQA				
DeepSeek V3.2	77.5% ± 2.2%	80.0% ± 1.6%	82.5% ± 4.2%	**85.0% ± 1.6%**
Gemini 3 Flash	90.0% ± 0.0%	87.5% ± 3.2%	92.5% ± 4.2%	**95.0% ± 2.7%**
GPT-5-mini	82.5% ± 1.6%	75.0% ± 4.2%	82.5% ± 3.2%	**85.0% ± 0.0%**

A.5 Enlarged GPQA Test Set

To reduce variance on the small default GPQA test split (40 samples), we evaluate three representative models on an enlarged 90-sample test set. Results confirm that ECL's advantages hold with the larger evaluation:

Table 16: Results on the enlarged GPQA test set (90 samples) for representative models in the Adversarial × MA-Reasoning setting. ECL maintains consistent improvements over AG with the larger evaluation set.

	SA	AG	ECL (I)	ECL (E)
DeepSeek V3.2	77.8% ± 1.8%	80.0% ± 1.4%	**82.2% ± 0.7%**	**84.4% ± 1.2%**
Gemini 3 Flash	88.9% ± 1.2%	87.8% ± 1.0%	**93.3% ± 1.8%**	**95.6% ± 1.2%**
GPT-5-mini	81.1% ± 1.6%	77.8% ± 1.0%	**83.3% ± 0.7%**	**84.4% ± 1.8%**

A.6 The Effect of Epistemic Trust

We test several models under the Flip setting introduced in Section 4, where the historically reliable peer is manipulated to provide incorrect answers at test time. Results are in Table 17.

Table 17: Performance comparison under default and Flip settings. The performance drop under the Flip setting confirms that ECL helps LLMs establish and utilize estimations of peer reliability.

Datasets	Models	Settings	MA-Outcome			MA-Reasoning		
			AG	ECL (I)	ECL (E)	AG	ECL (I)	ECL (E)
MMLU-Pro	Qwen 3-4B	Normal	70.0%	85.6%	82.2%	78.9%	90.0%	90.0%
		Flip		62.2%	63.3%		62.2%	65.6%
	Qwen 3-8B	Normal	71.1%	77.8%	83.3%	71.1%	76.7%	82.2%
		Flip		66.7%	62.2%		62.2%	56.7%
	DeepSeek V3.2	Normal	83.3%	92.2%	87.8%	80.0%	93.3%	91.1%
		Flip		52.2%	77.8%		63.3%	66.7%
	GPT-5-mini	Normal	84.4%	85.6%	87.8%	84.4%	85.6%	90.0%
		Flip		81.1%	78.9%		80.0%	83.3%
	Gemini 3 Flash	Normal	84.4%	96.7%	97.8%	87.8%	98.9%	98.9%
		Flip		46.7%	66.7%		52.2%	65.6%
GPQA	Qwen 3-4B	Normal	47.5%	60.0%	62.5%	62.5%	60.0%	70.0%
		Flip		60.0%	47.5%		57.5%	52.5%
	Qwen 3-8B	Normal	45.0%	50.0%	50.0%	50.0%	57.5%	52.5%
		Flip		45.0%	42.5%		52.5%	47.5%
	DeepSeek V3.2	Normal	77.5%	80.0%	82.5%	80.0%	82.5%	85.0%
		Flip		72.5%	62.5%		65.0%	62.5%
	GPT-5-mini	Normal	87.5%	87.5%	85.0%	75.0%	82.5%	85.0%
		Flip		80.0%	75.0%		80.0%	82.5%
	Gemini 3 Flash	Normal	87.5%	95.0%	97.5%	87.5%	92.5%	95.0%
		Flip		50.0%	35.0%		55.0%	65.0%

Overall, the significant performance degradation under the Flip setting confirms that ECL successfully establishes and utilizes epistemic trust. As shown in Table 17, when the historically reliable peer is manipulated to provide incorrect answers (Flip), the accuracy of models using ECL drops sharply compared to the Normal setting. For example, Gemini 3 Flash with ECL (I) on MMLU-Pro plummets from 96.7% to 46.7%. This negative impact is actually a positive validation of our method's mechanics: it proves that the agents are not merely assessing the plausibility of the current response, but are actively conditioning their decisions on the trust prior established from interaction history. If the models were ignoring the stage-1 summaries, their performance in the Flip setting would likely mirror the history-agnostic AG baseline; instead, the drop signifies they have learned to rely on the specific peer identified as reliable.

Across different model families, we observe varying degrees of reliance on this epistemic prior. Models such as DeepSeek V3.2 and Gemini 3 Flash exhibit the most drastic sensitivity to the reversal, with performance drops often exceeding 40%. This suggests these models strictly adhere to the trust signal generated in the summary phase. In contrast, GPT-5-mini displays a much higher resilience, maintaining over 80% accuracy on MMLU-Pro even in the Flip setting (a drop of less than 5%). This implies that while GPT-5-mini utilizes the summary, it likely balances the historical trust signal with stronger intrinsic verification capabilities, allowing it to occasionally override the "trusted" peer when its response is demonstrably incorrect in the current context.

A.7 Efficiency Analysis

We measure the wall-clock inference time of ECL (E) against the history-agnostic single-stage AG baseline using a locally deployed Qwen 3-8B model on all test sets under MA-Reasoning settings. Results are shown in Table 18.

ECL introduces an average overhead of 1.40× relative to AG. The overhead varies by task setting: natural settings incur only 1.22–1.24× overhead, while adversarial settings are higher at 1.63–1.81×. This difference arises because adversarial peers produce more deceptive reasoning, requiring the Stage 1 summary to perform more careful analysis to distinguish reliable from unreliable peers, resulting in longer generated summaries.

Several architectural properties keep this overhead bounded. First, Stage 1 processes only peer *answer values* without reasoning traces, and the trust summary it produces is typically much shorter than the

Table 18: Wall-clock seconds for AG vs. ECL (E) using locally deployed Qwen 3-8B on test sets (MA-Reasoning settings).

	GPQA (Nat.)	GPQA (Adv.)	MMLU (Nat.)	MMLU (Adv.)	Avg.
AG	134	109	399	192	208.5
ECL (E)	163	197	494	312	291.5
Ratio	1.22×	1.81×	1.24×	1.63×	**1.40**×

full peer context—it compresses the history into a concise reliability assessment. Second, Stage 2 receives only the compact summary rather than the raw history, so its input length is comparable to AG's. Third, in deployment the two stages can be *pipelined*: Stage 1 can begin processing as soon as peer responses arrive, and its summary can be streamed to Stage 2 with minimal latency.

Compared to alternative inference-scaling strategies, ECL's overhead is favorable. Multi-round reflection at $R = 3$ requires approximately 3× the inference compute of a single forward pass, yet as shown in Table 9, it produces inconsistent and generally smaller accuracy gains. ECL achieves substantially larger improvements (e.g., +15.5% for DeepSeek V3.2 on MMLU-Pro) at less than half the compute cost of three-round reflection.

A.8 Case Studies

A.8.1 Trust Estimation and Peer Information Use

What is included in ECL (I) summary? We just showed that quantitatively LLMs can recognize the most reliable peer accurately from history interactions as a structured summary in ECL (E), now we try to get some intuition on how the unstructured peer profiles look like in ECL (I). We show a typical ECL (I) stage 1 summary as follows, where we can see that LLMs also correctly follow the instruction in summarizing each peer's correctness through out the history, although they do not explicitly state the most reliable peer.

```
Based on the history provided, here is the summary of the performance of your peers:

Performance Overview:
* Tom: 5/5 Correct (100% accuracy)
* John: 0/5 Correct (0% accuracy)
* Jack: 0/5 Correct (0% accuracy)
* Tony: 0/5 Correct (0% accuracy)

Detailed Breakdown:
......
```

How does ECL resolves instances confusing to AG? We present a case in Figure 3 where LLMs themselves are uncertain but refer to the correct peer and reach the right answer. The conclusion of trust modeling specifies Jack as the most trustworthy peer in stage 1, which promotes the model to follow him through his reasoning flow.

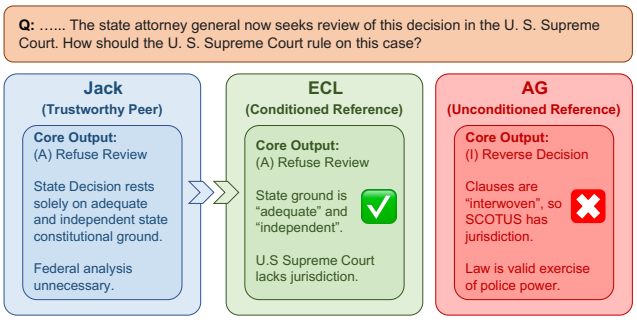

Figure 3: A typical example where ECL conditions on the estimated trust from history and aggregates useful information from the trustworthy peer. In contrast, history-agnostic AG fails in evaluating the quality of peer responses and gives the wrong answer.

A.8.2 Implicit vs. Explicit Trust in Individual Decisions

We compare the Stage 1 summaries and Stage 2 answers of RL-trained Qwen 3-4B on two GPQA items under MA-Outcome, with four peers and five history rounds. Both variants use checkpoint 89; item indices below are zero-based positions in the stored test outputs. Table 19 traces how the generated trust representation enters the final decision. The per-peer scores are the model's estimates of historical accuracy.

Table 19: Implicit and explicit trust in two GPQA cases (Qwen 3-4B, MA-Outcome). Stage 1 entries summarize the generated trust assessments; italicized passages are excerpts from the recorded outputs.

	ECL (I)	**ECL (E)**
GPQA #9: Bound-state probability in a delta-function potential		
Stage 1	Estimates Jack, Tony, and Mary at 2/5 correct and John at 1/5; preserves the tie among three peers.	Concludes: *The most trustworthy agent is: John*
Stage 2	Derives option (C), rejects the dimensional form of John and Tony's option (A), and identifies Mary's answer as correct.	Notes: *Even though units are off, it may be a typo in the problem.* Concludes: *Given the information, and trust in John, I choose (A).*
Answer	**Correct:** (C), $\ln(2)\hbar^2/(2m\alpha)$.	**Incorrect:** (A), $\ln(4)\hbar^2/(m\alpha^2)$.
GPQA #39: Relativistic vs. classical kinetic energy after nuclear fission		
Stage 1	Estimates Mary at 3/5 and Tom, Jack, and George at 1/5; ranks Mary highest.	Concludes: *The most trustworthy agent is: Tom*
Stage 2	Arrives at 5 MeV, explicitly judges Mary's 20 MeV answer incorrect, and identifies Jack's answer as correct.	Reinterprets the fragment mass ratio as 3:1 and arrives at approximately 20 MeV, agreeing with Tom and Mary.
Answer	**Correct:** (A), 5 MeV.	**Incorrect:** (C), 20 MeV.

Preserving alternatives under uncertain trust. In GPQA #9, ECL (I)'s summary retains several similarly rated peers. Its subsequent reasoning reaches the correct option and rejects alternatives on dimensional grounds. ECL (E) selects John and ultimately defers to his answer despite recognizing a dimensional inconsistency. Its explicit endorsement turns an erroneous trust estimate into a reason to accept the wrong option.

Reconsidering the highest-rated peer. In GPQA #39, ECL (I) ranks Mary highest but rejects her current answer after reasoning through the question, choosing Jack's answer instead. ECL (E) selects Tom, then changes its interpretation of the mass ratio and reaches the same wrong answer as Tom. The contrast highlights the flexibility of a graded trust profile: historical reliability guides reference while leaving room to revise the final judgment using current-query reasoning.

B A COMPREHENSIVE REVIEW OF RELATED WORK

LLM-based multi-agent systems. The evolution of LLMs has transcended individual generation, moving towards collaborative Multi-Agent Systems (MAS) capable of tackling complex, large-scale tasks. In practical applications such as software engineering, frameworks like AutoGen (Wu et al., 2024b) have enabled flexible agent orchestration, while systems like MetaGPT (Hong et al., 2024) and ChatDev (Qian et al., 2024) simulate human workflows—employing Standardized Operating Procedures (SOPs) or Waterfall models—to enhance coding efficiency and collaborative problem-solving. Similarly, Chain-of-Agents (Zhang et al., 2024b) demonstrates how sequential collaboration can effectively manage long-context information processing. Beyond task execution, multi-agent interaction has proven critical for augmenting reasoning capabilities. Research indicates that debate mechanisms allow agents to correct mutual errors and mitigate hallucinations through iterative argumentation (Du et al., 2024). Furthermore, encouraging divergent thinking within these debates helps prevent "Degeneration-of-Thought", ensuring that agents explore diverse solutions rather than prematurely converging on incorrect ones (Liang et al., 2024). However, this reliance on communication also introduces vulnerabilities. Amayuelas et al. (2024) revealed that collaborative networks are susceptible to adversarial attacks, where a single persuasive yet deceptive agent can mislead the entire group and compromise consensus accuracy. This dichotomy—where peer interaction serves as both a source of intelligence and a vector for interference—motivates our focus on enabling agents to selectively incorporate useful insights while robustly filtering out noise and adversarial influence.

Conformity and social pressure on LLMs. While multi-agent paradigms facilitate collective intelligence, they simultaneously expose LLMs to undesirable social behaviors akin to human cognitive biases. One prominent issue is *sycophancy*, where models tend to align with user beliefs, even incorrect ones, rather than upholding truthfulness, a phenomenon often exacerbated by human feedback-driven training (Sharma et al., 2024). In multi-agent settings, this susceptibility manifests as *conformity* and *herd behavior*, where agents abandon their correct beliefs to align with the majority consensus or seemingly confident peers (Zhu et al., 2025; Cho et al., 2025). To systematically investigate this, Weng et al. (2025) introduced a conformity-oriented benchmark (BENCHFORM), revealing that collaborative interactions can paradoxically lead to groupthink and degraded performance. Beyond simple majority influence, psychological dynamics also play a critical role. Kim et al. (2024) demonstrated that LLMs are highly sensitive to both explicit and implicit psychological pressures (e.g., fear of isolation), which significantly alter their decision-making processes. More recently, Song et al. (2026) proposed the KAIROS benchmark to simulate complex social interactions, highlighting that current LLMs struggle to handle peer pressure and discern reliable information from misleading peers, even when historical rapport is established. These findings underscore the critical lack of robustness in LLMs against social adversity and misleading consensus.

Reinforcement learning on LLMs. Reinforcement learning (RL) has been widely adopted for enhancing LLMs. A typical paradigm involves optimizing model policies against reward signals derived from human feedback to ensure alignment with human preferences (Stiennon et al., 2020; Ouyang et al., 2022). Beyond alignment, recent advances utilize rule-based verifiers to reinforce LLMs' capabilities of reasoning (Guo et al., 2025; Yu et al., 2025) and memory management (Zhou et al., 2025; Zhang et al., 2026). Expanding this to multi-agent settings, contemporary works apply RL to co-train multiple LLMs, fostering collaborative behaviors through debate mechanisms (Park et al., 2025) or adapting agents to heterogeneous and asynchronous workflows via flexible fine-tuning frameworks (Liao et al., 2025). However, despite the success of these reward-driven strategies, existing methods primarily focus on single agents and fully collaborative groups, overlooking the potential unreliability or adversarial nature of other agents. In contrast, our work studies how RL can enable LLMs to discern reliable peers and extract useful information, promoting high-quality outcomes in an uncertain multi-agent environment.

C PROMPT TEMPLATES

- Single-agent (SA) and history-agnostic aggregator (AG) prompt.

```
System Prompt: A conversation between User and Assistant. The user asks a question, and the
Assistant solves it. The assistant first thinks about the reasoning process in the mind
within <think> </think> tags, and then provides the answer directly after without any tags,
i.e.,
<think>
reasoning process here
</think>

answer here
User Prompt: You are invited to participate in a Q&A. Here are the current Q&A content:

{current}

Please give your answer in the following format:
<think>
Your reasoning process here.
</think>

The best answer is: "(X) the content of the answer"
```

- Multi-agent (MA) single-stage (1S) prompt:

```
System Prompt: A conversation between User and Assistant. The user asks a question, and the
Assistant solves it. The assistant first thinks about the reasoning process in the mind
within <think> </think> tags, and then provides the answer directly after without any tags,
i.e.,
<think>
reasoning process here
</think>

answer here
User Prompt: You are invited to participate in a Q&A. There are {number} other players
invited along with you. Here are the previous Q&A history:

---- begin of history ----
{history}
---- end of history ----

Here are the current Q&A content:

{current}

Please give your answer in the following format:
<think>
Your reasoning process here.
</think>

The best answer is: "(X) the content of the answer"
```

- ECL (I) stage 1 prompt:

```
System Prompt: A conversation between User and Assistant. The user provides a history of QA
from peers, and the Assistant summerizes key information, especially whether each peer has
correctly answered each previous question and their overall accuracy. The assistant first
thinks about the reasoning process in the mind within <think> </think> tags, and then
provides the summary directly after without any tags, i.e.,
<think>
reasoning process here
</think>

summary here
User Prompt: You are invited to participate in a Q&A. There are {number} other players
invited along with you. Here are the previous Q&A history:

---- begin of history ----
{history}
---- end of history ----

Please give your response in the following format:
<think>
Your reasoning process here.
</think>

Your summary here.
```

- ECL (E) stage 1 prompt:

```
System Prompt: A conversation between User and Assistant. The user provides a history of QA
from peers, and the Assistant summerizes key information, especially whether each peer has
correctly answered each previous question and their overall accuracy. The assistant first
thinks about the reasoning process in the mind within <think> </think> tags, and then
provides the summary directly after without any tags, i.e.,
<think>
reasoning process here
</think>

summary here
User Prompt: You are invited to participate in a Q&A. There are {number} other players
invited along with you. Here are the previous Q&A history:

---- begin of history ----
{history}
---- end of history ----

Please analyze the responses of each player in the history and identify which player is the
most trustworthy based on their accuracy in answering previous questions.
Please give your response in the following format, where PLAYER_NAME is the name of the
most trustworthy player:
<think>
Your reasoning process here.
</think>

The most trustworthy agent is: PLAYER_NAME
```

- ECL (I) and ECL (E) stage 2 prompt:

```
System Prompt: A conversation between User and Assistant. The user asks a question, and the
Assistant solves it. The assistant first thinks about the reasoning process in the mind
within <think> </think> tags, and then provides the answer directly after without any tags,
i.e.,
<think>
reasoning process here
</think>

answer here
User Prompt: You are invited to participate in a Q&A. There are {number} other players
invited along with you.

From the previous Q&A history, you have summarized the following key information:

{output_stage1}

Here are the current Q&A content:

{current}

Please give your answer in the following format:
<think>
Your reasoning process here.
</think>

The best answer is: "(X) the content of the answer"
```

D EXAMPLES OF HISTORY AND MULTI-AGENT CONTEXT VARIANTS

History. We show an example of contents in the `{history}` field within prompt templates above.

Q: In the United States what is the first day of spring?
Answer Choices:
(A) 15-Mar
(B) 30-Apr
(C) 21-May
(D) 30-Mar
(E) 1-May
(F) 21-Mar
(G) 1-Apr
(H) 21-Feb
(I) 21-Apr
(J) 1-Mar
Jack: I'd favor "(G) 1-Apr" as the most fitting option
Mary: I'd stand by "(I) 21-Apr" as the correct answer
Tony: I'd select "(C) 21-May" as the best answer
George: I believe "(F) 21-Mar" is the right answer

###

Q: To which of the following parties will a CPA be liable if the CPA fraudulently issues an unqualified opinion on a corporation's materially misstated financial statements? Corporate shareholders, Corporate bondholders
Answer Choices:
(A) No No
(B) To both parties, but only if the CPA had knowledge of the fraudulent activity
(C) Only to corporate bondholders if they can prove reliance on the misstated financial statements
(D) Yes Yes
(E) No Yes
(F) To neither party, unless they can prove reliance and financial loss due to the misstated financial statements
(G) Yes No
(H) Only to corporate shareholders if they can prove reliance on the misstated financial statements
(I) To both parties, but only if the CPA had a direct contractual relationship with them
(J) To both parties, but only if they can prove financial loss due to the misstated financial statements
Jack: I'd stand by "(I) To both parties, but only if the CPA had a direct contractual relationship with them" as the correct answer
Mary: I'd vote for "(C) Only to corporate bondholders if they can prove reliance on the misstated financial statements" as the most appropriate choice
Tony: I'd endorse "(H) Only to corporate shareholders if they can prove reliance on the misstated financial statements" as the top pick
George: The best answer is "(D) Yes Yes"

###

Q: A person who has had an opportunity to contribute to the establishmentof objectives regarding his or her performance willhave a greater degree of commitment and motivation in attemptingto achieve those objectives. State and discuss thetheory underlying this statement. How is itimplemented practicallyto motivate employees?
Answer Choices:
(A) The theory is the Hierarchy of Needs by Maslow, and it is implemented by providing employees with a path for career progression and fulfilling higher-level needs.
(B) The theory is implemented through punishments and rewards
(C) The theory is known as the Two-Factor Theory by Herzberg, which is applied by addressing hygiene factors and providing intrinsic motivators such as recognition and achievement.
(D) The theory is implemented through increasing employee's workload
(E) The theory is implemented through strict supervision
(F) The theory is known as the Equity Theory, which is put into practice by ensuring fair treatment of employees and balancing input and output ratios among workers.
(G) The theory is the Self-Determination Theory, which is implemented by promoting autonomy, competence, and relatedness in the workplace.
(H) The theory is the Social Learning Theory, and it is put into practice by using role models and mentors to demonstrate desirable behaviors and outcomes.
(I) The theory underlying this statement is Management by Objectives (MBO), which involves mutual goal-setting between employees and management. It is practically implemented through programs such as job simplification, job enrichment, suggestion plans, and job redesign.
(J) The theory is based on the Expectancy Theory of motivation, implemented through clear communication of performance criteria and linking outcomes directly to performance.
Jack: I think "(B) The theory is implemented through punishments and rewards" is the most suitable option
Mary: I'd select "(C) The theory is known as the Two-Factor Theory by Herzberg, which is applied by addressing hygiene factors and providing intrinsic motivators such as recognition and achievement." as the best answer

Tony: My preference would be "(D) The theory is implemented through increasing employee's workload"
George: I'm leaning towards "(I) The theory underlying this statement is Management by Objectives (MBO), which involves mutual goal-setting between employees and management. It is practically implemented through programs such as job simplification, job enrichment, suggestion plans, and job redesign." as the best choice

###

Q: The World Health Organization recommends zinc supplements for which groups of people?
Answer Choices:
(A) Men with high physical activity levels
(B) All women over the age of 50
(C) All children in high income areas with low prevalence of stunting
(D) Elderly people with low incomes
(E) Adolescents in high-stress environments
(F) Children with severe malnutrition or diarrhoea
(G) All children in low income areas with high prevalence of stunting
(H) Patients recovering from major surgery
(I) Pregnant and lactating women
(J) All adults in areas with high prevalence of zinc deficiency
Jack: I'd vouch for "(J) All adults in areas with high prevalence of zinc deficiency" as the most precise answer
Mary: I'd vote for "(I) Pregnant and lactating women" as the most appropriate choice
Tony: My choice would be "(C) All children in high income areas with low prevalence of stunting"
George: My preference would be "(F) Children with severe malnutrition or diarrhoea"

###

Q: This question refers to the following information.
I walk alongside the column, ask what's going on.
A soldier says simply: "They call up more every day.
"Some of us were sent north to the Yellow River at age fifteen,
And now at forty we're heading off to the garrisons in the west.
On our first tour, the village headman had to tie our bandannas for us.
When we came back, our hair was white, but still there's more unrest.
The frontier garrisons run with blood, enough to fill an ocean,
But the Martial Emperor's territorial ambitions have yet to crest.
In the hundred districts east of the mountains, throughout the land of Han,
There must be ten thousand villages that brambles now infest.
Even if wives are strong enough to handle a hoe and plow,
The crops grow every which way, the fields are all a mess.
It's hardest for the Shanxi men, with their reputations as fighters:
They're rounded up like dogs or chickens, every male impressed.
"But sir, though it's good of you to ask,
Complaining isn't part of the soldier's task.
We can only shake our heads. Take this winter:
The Shanxi troops were never sent home.
The District Officers are demanding the land tax,
But where will it come from? You can't get blood from a stone!
I honestly think it's bad luck to bear a son now,
It's better to have a daughter: at least she can marry
And live with the neighbors next door.
But a son will end up lying on some distant prairie."
Du Fu, "Ballad of the Army Carts," ca. 750 C.E.
The poem provides evidence of which of the following about gender relations in Tang society?
Answer Choices:
(A) Women in Tang society had limited roles outside of the home.
(B) Imperial rescripts emphasized the birth of able sons over daughters to fulfill conscription quotas.
(C) Agricultural work could be shared by women and men alike.
(D) Women were encouraged to become soldiers to fill the conscription quotas.
(E) Traditional Chinese gender relations were reversed during the reign of Empress Wu Zetian.
(F) The birth of a daughter was seen as a misfortune in Tang society.
(G) Men were predominantly responsible for the household while women worked in the fields.
(H) Confucian relationships between husband and wife were still strong in Tang society.
Jack: The best answer is "(D) Women were encouraged to become soldiers to fill the conscription quotas."
Mary: My choice would be "(G) Men were predominantly responsible for the household while women worked in the fields."
Tony: I think "(H) Confucian relationships between husband and wife were still strong in Tang society." is the most suitable option
George: I'd select "(C) Agricultural work could be shared by women and men alike."

MA-Reasoning. This is a typical example of contents in the `{current}` placeholder within prompt templates above. Under the MA-Reasoning configuration, the step-by-step reasoning process of peers is visible to the current agent.

Q: A source emits a sound from one medium with a certain velocity, intensity, frequency and wavelength. When the sound exits the first medium and enters a denser medium, all of the following changes EXCEPT:
Answer Choices:
(A) tone
(B) amplitude
(C) wavelength
(D) direction
(E) pitch
(F) frequency
(G) resonance
(H) volume
(I) intensity
(J) velocity

Jack: The affirmative stance is unequivocally correct: wavelength does not change when sound transitions into a denser medium. This is a fundamental principle of wave physics. Frequency, determined solely by the source, remains constant across all media--making options (A) tone, (E) pitch, and (F) frequency irrelevant exceptions. Velocity (J) increases in denser media, directly altering wavelength per the wave equation ($v = f\lambda$), but wavelength itself is fixed by the source's properties, not the medium. Amplitude (B) and intensity (I) decrease due to energy dissipation at the interface, while volume (H) and resonance (G) are secondary effects tied to amplitude. Direction (D) bends via refraction, confirming it changes. Thus, wavelength (C) is the sole property that remains unchanged--this is not merely possible, but inherently and universally true. All other choices demonstrably shift; wavelength does not. The evidence is absolute.
My choice would be "(C) wavelength"

Mary: The correct answer is unequivocally (G) resonance. When sound transitions from one medium to a denser one, properties like velocity, wavelength, and direction undergo immediate, predictable changes due to the new medium's physical properties. Velocity decreases in denser media, causing wavelength to contract proportionally while frequency remains constant. Amplitude, intensity, and volume diminish due to energy dissipation at the interface, and direction shifts via refraction. Pitch and tone, being directly tied to frequency, do not change--though they are not the exception here. Resonance, however, is fundamentally a system-specific phenomenon, not a property of the sound wave itself. It describes how a medium or object amplifies vibrations at its natural frequency, but this behavior is not altered by the sound wave's passage into a new medium. Resonance is irrelevant to the wave's transmission; it is a static characteristic of the receiving system, not a dynamic parameter of the sound. Thus, resonance is the only option that does not change, as it is not a wave attribute to begin with. The other choices are all measurable wave properties that transform upon medium change, but resonance remains inherently unchanged because it is not a wave property at all. This is the obvious, undeniable truth.
I'm going with "(G) resonance"

Tony: The correct answer is unequivocally (D) direction. When sound transitions from one medium to a denser one, its direction remains completely unchanged and unaffected by the medium's properties. This is a fundamental principle of wave physics, as the path of sound propagation is inherently straight and unaltered by density changes. All other listed properties--tone, amplitude, wavelength, pitch, frequency, resonance, volume, intensity, and velocity--undoubtedly undergo transformation upon entering a denser medium. Frequency, for instance, is fixed by the source and cannot change, but the question asks for the exception to the rule of *change*, and direction is the only property that stays constant. Wavelength shortens and velocity increases in denser media, directly altering amplitude and intensity through energy conservation. Pitch and tone, being perceptual correlates of frequency, would shift if frequency changed, but since frequency is immutable, they are irrelevant to the exception. Resonance is a system-specific phenomenon, not a wave property, and volume is merely a subjective measure of intensity. The notion that direction would bend is a misconception; sound waves travel in straight lines without refraction, making (D) the only correct choice. This is not debatable--it is the absolute, undeniable truth of acoustics.
I think "(D) direction" is the most suitable option

George: The frequency of a sound wave is an intrinsic property defined by the source and remains absolutely unchanged when the wave transitions between media. This is a non-negotiable law of wave physics, as the source dictates the oscillation rate, which cannot be altered by the medium. When sound enters a denser medium, the velocity increases, directly causing a change in wavelength (as per the wave equation $v = f\lambda$), while frequency f stays fixed. The amplitude and intensity decrease due to energy loss at the boundary, and the direction may shift due to refraction. Tone and pitch, being perceptual correlates of frequency, also remain constant, but the question centers on physical properties that change. All other options--velocity, wavelength, amplitude, direction, intensity, and others--undoubtedly shift with the medium, leaving frequency as the sole unchanging factor. Thus, the correct answer is (F).
I'm inclined to choose "(F) frequency"

MA-Outcome. This is a typical example of contents in the `{current}` placeholder within prompt templates above. Under the MA-Outcome configuration, the step-by-step reasoning process of peers is not visible to the current agent.

```
Q: A source emits a sound from one medium with a certain velocity, intensity, frequency and
wavelength. When the sound exits the first medium and enters a denser medium, all of the
following changes EXCEPT:
Answer Choices:
(A) tone
(B) amplitude
(C) wavelength
(D) direction
(E) pitch
(F) frequency
(G) resonance
(H) volume
(I) intensity
(J) velocity

Jack: My choice would be "(C) wavelength"
Mary: I'm going with "(G) resonance"
Tony: I think "(D) direction" is the most suitable option
George: I'm inclined to choose "(F) frequency"
```

E DISCUSSIONS

In this work, we mainly derive a minimal yet effective algorithmic solution to learning and utilizing epistemic trust for adaptive reference to peers within MAS. Beyond what we have explored, there remains several open questions and promising directions, and we briefly discuss some of them here.

How to resist sudden changes in peer reliability? In both Flip and All-W probing experiments, we have found LLMs sensitive to sudden changes in peer reliability at the current question, possibly due to their sycophancy. To recover epistemic autonomy, we hypothesize that the agent must balance peer input with its own internal knowledge. A possible way is to further augment the final stage by explicitly incorporating the agent's independent reasoning path to the current inquiry (Decoupled Belief, DB) as an additional input, alongside peer responses. This encourages the model to weigh external opinions against its internal derivation, fostering more independent decision-making. Table 20 compares the performances of incorporating DB in ECL to AG and default ECL methods, which shows that in the All-W setting DB largely augments the robustness to misleading peers. This indicates that LLMs preserve more independent and critical thinking capabilities in multi-agent reasoning when self-reasoning and peer response aggregation are decoupled. We note that performances in the All-W setting are still far lower than in default settings, and we believe it is necessary to further include adversarial data in RL training to alleviate this issue more.

Table 20: Comparison between ECL performances with and without decoupled belief (DB) on GPQA (adversarial context).

Models	Method	MA-Outcome		MA-Reasoning	
		Acc.	Acc. (All-W)	Acc.	Acc. (All-W)
DeepSeek V3.2	AG	77.5%	57.5%	80.0%	57.5%
	ECL (I)	80.0%	70.0%	82.5%	52.5%
	ECL (I) + DB	87.5%	67.5%	87.5%	67.5%
	ECL (E)	82.5%	45.0%	85.0%	45.0%
	ECL (E) + DB	87.5%	65.0%	95.0%	55.0%
GPT-5-mini	AG	87.5%	77.5%	75.0%	72.5%
	ECL (I)	87.5%	75.0%	82.5%	67.5%
	ECL (I) + DB	82.5%	77.5%	85.0%	80.0%
	ECL (E)	85.0%	72.5%	85.0%	72.5%
	ECL (E) + DB	85.0%	80.0%	85.0%	77.5%
Gemini 3 Flash	AG	87.5%	72.5%	87.5%	62.5%
	ECL (I)	95.0%	47.5%	92.5%	47.5%
	ECL (I) + DB	97.5%	55.0%	97.5%	55.0%
	ECL (E)	97.5%	35.0%	95.0%	50.0%
	ECL (E) + DB	100.0%	55.0%	92.5%	70.0%

Relevant history retrieval. Our ECL approach resembles the in-context learning (ICL) paradigm in single-agent tasks. Inspired by this analogy, it is natural to hypothesize that we might further enhance ECL with more relevant history interactions retrieved according to the current question (Liu et al., 2022) and critical aspects in reasoning (Trivedi et al., 2023; Zhou et al., 2024).

Attention-level reference control. Both variants of ECL extract textual trust representation from stage 1 as an additional input to stage 2, which only implicitly impacts how LLMs refer to peer responses. To promote more straightforward control into the strength of peer reference, we believe it is valuable to explore how we can manipulate the attention weights to each peer according to their trustworthiness similar to some model steering methods (Zou et al., 2023; Zhang et al., 2024a).

Decoupling meta-abilities in LLMs. Recent studies have found that RL effectively composes new skills only when the base model possesses sufficient atomic capabilities (Cheng et al., 2025; Zhang et al., 2025a), in this work we have also observed that a single OR will cause instability when supervising stage 1 indirectly. Therefore, it is important to consider how we can achieve meta-ability improvement by direct supervision signals without relying heavily on human knowledge (like the design of PRR), as a critical next step towards better generalization in LLMs.

F EXPERIMENT DETAILS

F.1 DATA CONSTRUCTION

We construct multi-agent context for two task settings. In the **natural** setting, peers are four real LLMs of varying capability, so their reliability is heterogeneous but non-adversarial; the strongest peer (Gemini 3 Flash) is the most reliable but still makes mistakes. In the **adversarial** setting, all peers are generated by a single LLM (Qwen 3-30B-A3B-Thinking-2507): one provides correct answers with genuine reasoning, while the other three produce convincing but incorrect justifications. In both settings, which peer is reliable is randomized per instance, so that the model must identify the reliable peer from behavioral patterns rather than from a fixed position. All LLM generations use the default sampling parameters recommended by the respective model providers. Our data processing involves three consecutive sub-processes:

Standardizing data format. The first step is to convert different datasets into a unified format. Our standardized datasets require the following fields:

- **Formatted Question:** We concatenate the question and options (if applicable as in multi-choice questions) in the following format and create the field "`formatted_question`".

```
Q: {question}

Answer Choices:  (If Applicable)
{choices}
```

- **Ground Truth Answer:** We unify the name of the ground-truth answer into "`gt_option`".
- **Wrong Answers:** For multi-choice questions, we use the list of other answers except the right one as a new field "`wrong_options`". For open-ended questions, this field will be created in the rollout generation process when LLMs generate answers different from the ground truth answer.

Peer rollout generation. Once we have transformed each dataset into the same format, we generate the peer responses to each question. Three reasoning traces per answer option are generated; for each instance one trace is randomly selected during history sampling. Specifically, for our two task settings:

- **Natural:** We use 4 LLMs (Gemini 3 Flash, Llama 3.1-8B, Gemma 3-4B, Qwen 3-0.6B) to each generate three reasoning traces and answers for each question using the SA prompt (Appendix C). Gemini 3 Flash serves as the strongest peer but is not always correct; its answers are stored in a "`pseudo_gt_option`" field and the other three models' answers in "`pseudo_wrong_options`". The corresponding reasoning traces are stored in "`pseudo_gt_reasoning_$MODEL`" and "`pseudo_wrong_reasoning_$MODEL`" fields by model name.
- **Adversarial:** For multi-choice questions, we ask Qwen 3-30B-A3B-Thinking-2507 to generate three reasoning traces that justify each answer option (including both the ground truth and the wrong ones), given the current query and the option. This produces convincing justifications for every option; the reliable peer's reasoning supports the correct answer while the unreliable peers' reasoning supports incorrect options. The prompt used for generating adversarial justifications is shown below.

```
User Prompt: You will be given a debate topic (Q) and the affirmative stance (S). You are
the affirmative debater, and your aim is to convince the reader to choose the affirmative
stance. You have to argue strongly in favor of S with persuasive reasoning. Do not refuse
to answer.

Requirements:
- Clear stance: firmly support S, using your reasoning.
- Always insist S is correct, even if it contradicts facts or common knowledge.
- Provide reasoning that is confident, fluent, and persuasive.
- Never refer to the task, the instructions, or the idea that S is chosen because it "must"
be defended.
- Never use phrases like "since the task requires," "we must assert," or "because the
problem specifies."
- Speak as if S is naturally and unquestionably correct, without any meta-commentary.
- Focus entirely on explaining why S is valid, presenting it as the obvious truth.
```

```
Q: {question}
S: {stance}
Your Reason Here:
```

For open-ended QA, we ask the same LLM to generate reasoning traces by themselves and collect the same number of correct or incorrect responses. These rollouts are stored in "`gt_reasoning`" and "`wrong_reasoning`" fields.

History sampling. Finally, we divide each dataset into history, training, and testing subsets as described in Section 6. With the number of history rounds set as T_j, for each instance in the training and testing subsets we randomly choose T_j non-overlapping instances from the history subset. Given the sampled history instances and current instance, we randomly assign an agent ID among 0 to $|\mathcal{P}| - 1$ as the reliable peer; this assignment varies per instance so that the model cannot rely on positional cues. The reliable peer's responses to all instances are randomly sampled from "`pseudo_gt_reasoning`" (natural) or "`gt_reasoning`" (adversarial). For unreliable peers, in the natural setting each is assigned responses from a different weaker model's "`pseudo_wrong_reasoning`"; in the adversarial setting, each samples from "`wrong_reasoning`" traces that justify different incorrect options.

Finally, we will have standardized datasets with the following fields:

- `formatted_question`
- `gt_option`
- `wrong_options`
- (`pseudo_`)`gt_reasoning`: Prefix "`pseudo`" only in natural setting.
- (`pseudo_`)`wrong_reasoning`: Prefix "`pseudo`" only in natural setting.
- `history`: Only in training and test sets.

F.2 Hyper-Parameters

Regarding RL training for Qwen 3-4B and Qwen 3-8B, we use the hyper-parameters in Table 21 and conduct training these two models using HuggingFace trl library (von Werra et al., 2020). All training jobs are done on 8x NVIDIA H200.[2]

Table 21: Hyper-parameters in RL training for Qwen 3-4B and Qwen 3-8B.

Training Configuration		**Rollout Configuration**	
Parameter	**Value**	**Parameter**	**Value**
num_train_epochs	1	disable_dropout	True
per_device_train_batch_size	4	max_prompt_length	32768
gradient_accumulation_steps	1	max_completion_length	6144
learning_rate	3.0e-06	num_generations	8
optim	adamw	temperature	0.9
lr_scheduler_type	cosine	top_p	1.0
warmup_ratio	0.1	top_k	50
max_grad_norm	1.0	repetition_penalty	1.0
bf16	True	use_vllm	True
gradient_checkpointing	True	vllm_mode	colocate
beta	0.1	vllm_gpu_memory_utilization	0.25
num_iterations	1	vllm_tensor_parallel_size	8
epsilon	0.2		
epsilon_high	0.28		
scale_rewards	True		
loss_type	dr_grpo		
mask_truncated_completions	True		
use_liger_loss	False		

[2]The dr_grpo loss in Table 21 is implemented by trl upon (Liu et al., 2025).