RPO-MAS: RECEIVER PREFERENCE OPTIMIZATION FOR SELECTIVE MULTI-AGENT COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication is a critical component of multi-agent reasoning, yet peer communication does not always improve the receiver's decision. A peer response that is correct in isolation may be redundant or harmful for a receiver with a different reasoning state, while autonomous refinement or preserving the current response may lead to better outcomes. Existing approaches for selective multi-agent communication mainly rely on source reliability, peer selection, or communication structures, providing limited supervision on whether a particular intervention benefits the receiver. In this work, we propose RPO-MAS, a receiver-conditioned preference optimization framework for selective multi-agent communication that learns intervention preferences from receiver-side outcomes. RPO-MAS formulates communication selection over three intervention types: peer-conditioned updating (SEND), autonomous refinement (SELF), and state preservation (KEEP). Starting from the same receiver state, it compares alternative interventions and constructs outcome-based preferences from their observed correctness outcomes. Based on these preferences, RPO-MAS learns a receiver-conditioned scorer to select appropriate interventions among candidate actions. Evaluation across seven reasoning benchmarks with Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct agents shows that RPO-MAS consistently improves multi-agent reasoning performance over independent aggregation and communication-based baselines. Compared with Majority Voting, RPO-MAS improves average accuracy from 76.91% to 77.72% and from 72.42% to 73.13% on the two backbones, respectively. Further analysis shows that receiver-conditioned supervision improves correction of erroneous states while reducing harmful changes to correct states.

1 INTRODUCTION

A correct language agent can enter a debate and leave with the wrong answer. While multi-agent communication is intuitively designed to enable mutual error correction (Du et al., 2024; Liu et al., 2025), extensive empirical evidence reveals a troubling counter-phenomenon: unstructured debate often amplifies errors, induces sycophantic conformity, and degrades into consensus around incorrect solutions (Weng et al., 2025; Zhu et al., 2025; Cui et al., 2026). Indeed, simple majority voting over independent reasoning steps frequently outperforms multi-round deliberation (Choi et al., 2025). More opportunities to communicate therefore do not reliably produce better decisions; they can instead create additional opportunities for a correct reasoning state to be displaced by misleading peer information (Kaesberg et al., 2025; Fan et al., 2026).

One unresolved bottleneck is that the criterion used to select communication need not coincide with the benefit received by a particular agent: **message quality and receiver benefit are distinct quantities.** Recent systems already make communication selective by estimating source reliability, pruning redundant communication structures or viewpoints, and dynamically choosing which peers to reference (Wang et al., 2025; Lin & Hooi, 2025; Zhang et al., 2025; Zeng et al., 2025; Wu et al., 2026). However, these selection signals do not directly measure the consequence of applying a particular message to a particular receiver from its current state. A factually sound argument is functionally redundant to an agent that already possesses the correct rationale, unconvincing to one anchored in a conflicting misconception, and potentially disruptive to a fragile yet correct receiver state. A message that is imperfect in isolation can still benefit a receiver when it addresses a receiver-specific error.

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.

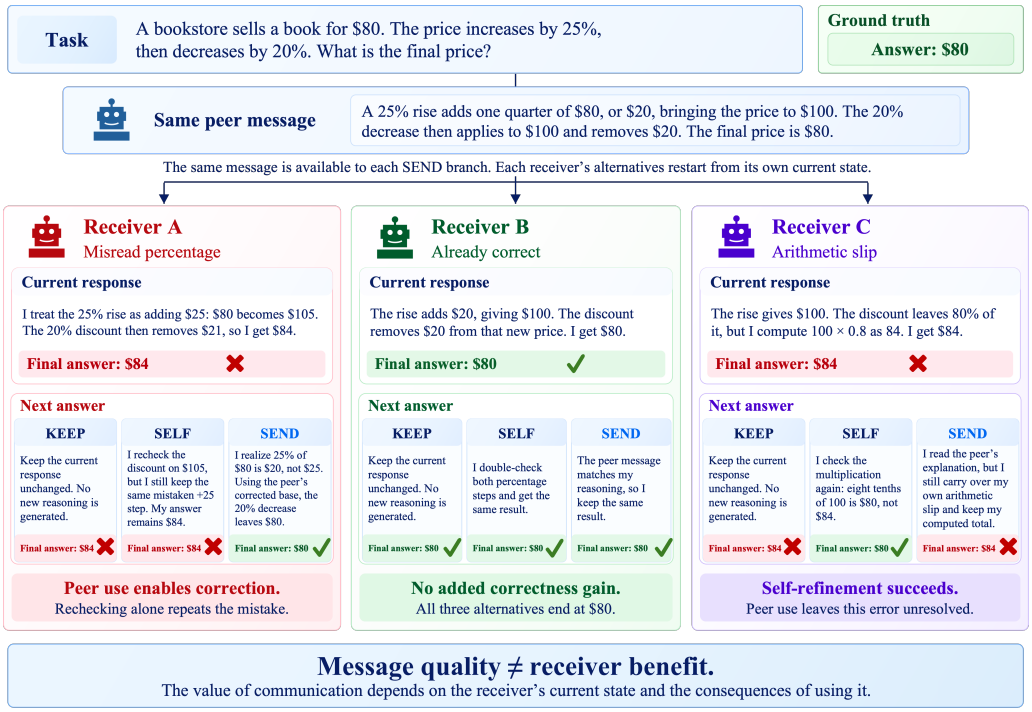

Figure 1: The Same Message Has Different Value under Different Receiver States.

Therefore, communication selection should consider receiver-side outcomes rather than message quality alone, as the same message can have different values across receiver states.

To formalize this decision, an agent must distinguish the effect of peer information from autonomous refinement and state preservation. At any reasoning stage, an autonomous receiver confronts three distinct intervention types: (1) Incorporate external peer reasoning (SEND), (2) Refine reasoning autonomously (SELF), or (3) Preserve the current state without perturbation (KEEP) to prevent self-doubt and unnecessary computation (Huang et al., 2024; Chen et al., 2025b). Evaluating communication requires a counterfactual baseline that contrasts peer-directed updates against both autonomous refinement and state preservation. This minimal intervention set separates external information from additional computation and non-intervention, exposing two complementary receiver-side objectives: Opportunity to Correct: steering an erroneous receiver toward the correct answer through beneficial peer exchange or independent refinement (Liu et al., 2025; Kumar et al., 2025). Opportunity to Protect: insulating an already successful receiver state from disruptive external misguidance or overthinking-induced degradation (Figure 1).

In this work, we translate this counterfactual formulation into a direct learning signal, introducing Receiver Preference Optimization for Multi-Agent Systems (RPO-MAS). Given a fixed receiver state, RPO-MAS executes candidate peer-conditioned branches (SEND) alongside independent self-refinement (SELF) under identical decoding conditions, while evaluating state preservation (KEEP) directly on the receiver's current answer. Rather than relying on source-side or structural proxies, we supervise action preferences directly from the correctness outcomes induced at the receiver. Actions with equal observed outcomes remain unordered, while strict outcome differences define pairwise preferences over the available interventions. RPO-MAS trains a lightweight receiver-conditioned scorer with a group-normalized pairwise ranking objective and directional reweighting that mitigates imbalance in Update–Keep and Peer–Self comparisons. At inference time, the scorer evaluates each candidate intervention from the receiver's observable reasoning state and decides whether to use a particular peer, refine independently, or preserve the current response.

Our contributions are threefold:

- We formulate multi-agent communication as a receiver-conditioned intervention selection problem that separates peer information, autonomous refinement, and state preservation through three actions: SEND, SELF, and KEEP.
- We propose RPO-MAS, which constructs receiver-conditioned intervention preferences from matched receiver-side execution outcomes rather than static message quality. We learn a preference scorer to select interventions based on receiver states.
- We evaluate RPO-MAS across multiple reasoning benchmarks and model families, with controlled ablations isolating receiver conditioning, matched execution supervision, and directional balancing.

2 Related Work

2.1 Multi-Agent Reasoning: Deliberation vs. Independent Aggregation

Exchanging reasoning steps through multi-agent debate allows models to correct mistakes (Du et al., 2024). However, recent studies show that simple majority voting over independent responses often matches or even outperforms repeated debate (Choi et al., 2025). Unconstrained communication can also cause agents to converge on incorrect answers and propagate errors across the system (Weng et al., 2025; Tian et al., 2026). More importantly, a debate update changes two factors at once: the receiver obtains external peer information and also receives an additional opportunity to reason. Since continued self-refinement can itself alter or improve an answer (Madaan et al., 2023), observing a successful post-communication update does not determine whether the improvement comes from the peer message itself. Standard debate-versus-voting comparisons therefore do not separate the value of external information from the effects of additional computation or non-intervention. RPO-MAS makes these alternatives explicit by evaluating peer-conditioned updating (SEND), autonomous self-refinement (SELF), and state preservation (KEEP) from the same receiver checkpoint, using SELF as an active-computation control and KEEP as a non-intervention control.

2.2 Selective Communication: From Message Selection to Receiver Consequences

To reduce unnecessary or harmful communication, prior work has investigated communication topology design and edge pruning (Zhang et al., 2025), adaptive debate triggering (Fan et al., 2026), and filtering unreliable information from previous interactions (Tian et al., 2026). Other approaches estimate peer reliability from interaction history (Zhou et al., 2026) or dynamically determine which peers an agent should consult (Wu et al., 2026). Together, these methods show that communication can be selectively controlled. However, their selection signals do not directly capture the receiver-side consequences of applying a particular message. A message produced by a reliable peer may be redundant for a receiver that is already correct, while the same message may be useful for another receiver whose current reasoning contains a compatible error. Conversely, a generally plausible message can still disrupt an otherwise correct receiver state. RPO-MAS therefore defines intervention preferences through receiver-side outcomes: from a fixed receiver state, it executes candidate peer-conditioned interventions and compares their consequences with autonomous self-refinement and state preservation.

2.3 Preference Optimization for Interaction Policies

Preference-based learning provides a natural mechanism for converting comparative feedback into decision policies. DPO learns directly from offline preference pairs without requiring an explicit reward model or an on-policy reinforcement-learning loop (Rafailov et al., 2023), whereas GRPO is a reinforcement-learning method that estimates relative advantages from groups of sampled responses (Shao et al., 2024). In multi-agent settings, OPTIMA constructs preference data from interaction trajectories using task-level rewards (Chen et al., 2025a), while MALT expands multi-agent reasoning trajectories through tree search and assigns credit to intermediate agent decisions from joint outcomes (Motwani et al., 2025). These methods demonstrate how trajectory-level feedback and credit assignment can be used to improve multi-agent behavior. RPO-MAS differs in the granularity at which supervision is constructed. Rather than inferring the value of a local communication decision from a longer interaction trajectory, it fixes the receiver checkpoint and directly executes the

mutually exclusive interventions available from that state. Strict outcome differences induce pairwise preferences, while equal-outcome actions remain unordered; a lightweight receiver-conditioned scorer is then trained to select among SEND, SELF, and KEEP.

3 PROBLEM FORMULATION

3.1 MULTI-AGENT INTERACTION

Consider a team of N language-model agents solving a question x with ground-truth answer y over interaction rounds $t \in \{0, \dots, T-1\}$. At the beginning of round t, agent i has produced a response R_i^t with parsed answer $\hat{y}_i^t = \mathrm{ans}(R_i^t)$. We denote its observable state by $s_i^t = (R_i^t, \hat{y}_i^t)$ and the public snapshot by $\mathcal{S}^t = \{s_1^t, \dots, s_N^t\}$. For receiver i, the response of another agent j is a candidate peer message $m_j^t = R_j^t$. The corresponding decision context is $z_i^t = (x, s_i^t, \{m_j^t\}_{j \neq i})$.

3.2 RECEIVER-LEVEL INTERVENTION DECISION

From its current state, receiver i may preserve the response, continue reasoning independently, or condition its next response on a peer message. We write the intervention set as

$$\mathcal{A}_i^t = \{\mathrm{K}, \mathrm{S}\} \cup \{\mathrm{M}_j : j \neq i\}, \tag{1}$$

where K denotes KEEP, S denotes SELF, and M_j denotes SEND(j). The three intervention types correspond respectively to state preservation, an additional autonomous reasoning step, and an additional reasoning step conditioned on peer j.

Let F denote the underlying language model and ξ its generation randomness. Each action in $\mathcal{A}_i^t$ defines an alternative continuation from the same receiver state:

$$\widehat{R}_i^{t+1}(a; \xi) = \begin{cases} R_i^t, & a = \mathrm{K}, \\ F(x, R_i^t; \xi), & a = \mathrm{S}, \\ F(x, R_i^t, m_j^t; \xi), & a = \mathrm{M}_j. \end{cases} \tag{2}$$

The KEEP action leaves the current response unchanged, whereas SELF and SEND each produce one new receiver response.

We evaluate an intervention by the correctness of the final answer it induces at the receiver. Its expected utility is

$$U(a \mid z_i^t) = \Pr_{\xi}\left[\mathrm{ans}\left(\widehat{R}_i^{t+1}(a; \xi)\right) = y\right]. \tag{3}$$

The receiver-level decision is therefore

$$a_i^{t,*} \in \arg\max_{a \in \mathcal{A}_i^t} U(a \mid z_i^t). \tag{4}$$

For two interventions $a, b \in \mathcal{A}_i^t$, we write $a \succ_{z_i^t} b$ when $U(a \mid z_i^t) > U(b \mid z_i^t)$; equal utilities define a tie. Since z_i^t includes the receiver's current response, the ordering over the same candidate interventions may vary across receiver contexts.

A peer-conditioned continuation combines an additional reasoning step with external information. The SELF action provides the corresponding continuation without peer conditioning, while KEEP provides the non-intervention reference. Their utilities satisfy

$$U(\mathrm{M}_j \mid z_i^t) - U(\mathrm{K} \mid z_i^t) = \underbrace{U(\mathrm{S} \mid z_i^t) - U(\mathrm{K} \mid z_i^t)}_{\text{autonomous deliberation}} + \underbrace{U(\mathrm{M}_j \mid z_i^t) - U(\mathrm{S} \mid z_i^t)}_{\text{incremental peer-conditioned value}}. \tag{5}$$

The first term measures the effect of an additional autonomous generation, while the second measures the incremental value associated with conditioning on peer j.

4 RPO-MAS

RPO-MAS learns a receiver-conditioned scorer over the intervention set in Eq. (1). As illustrated in Figure 2, RPO-MAS evaluates alternative interventions from the same receiver state and learns

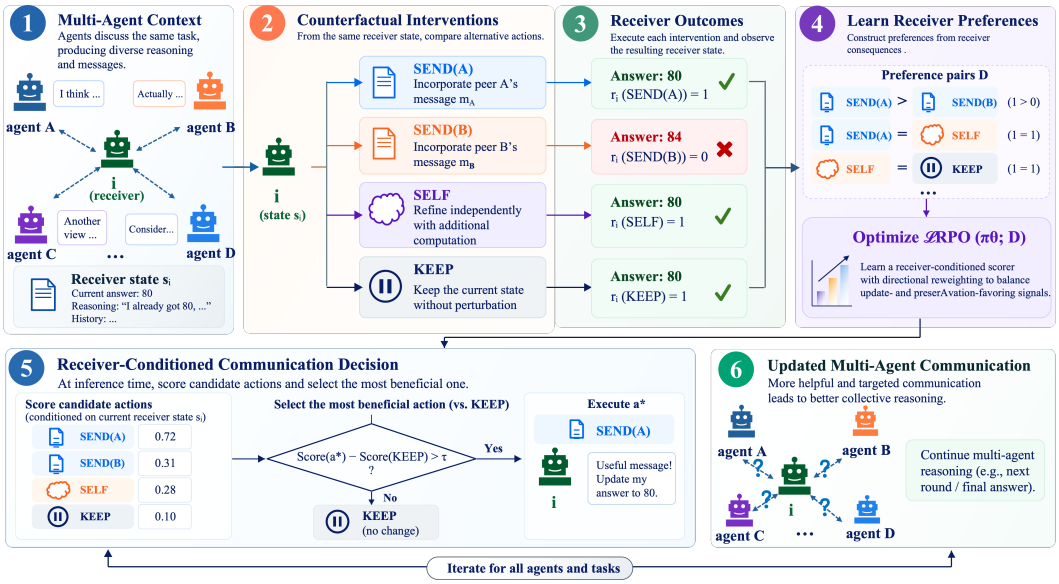

Figure 2: Overview of RPO-MAS. Starting from the same receiver state, RPO-MAS evaluates peer communication, self refinement, and state preservation, then learns preferences over these actions based on their observed outcomes.

preferences from their observed outcomes. For each training receiver state, alternative interventions are executed from the same checkpoint and evaluated by their receiver-side outcomes. Strict outcome differences define intervention preferences over the available actions, which are then used to train a scorer $f_\theta(z_i^t, a)$. At inference time, the scorer ranks the available interventions before execution and selects the intervention to apply for the current receiver state.

4.1 Matched Outcome Supervision

For a training receiver i with context z_i^t, we evaluate every available intervention from the same pre-action state s_i^t. The receiver state and candidate messages are fixed across the alternative branches, and all stochastic update branches use the same decoding seed ξ_i. For $a \in \mathcal{A}_i^t$, the realized receiver outcome is

$$r_i(a; \xi_i) = \mathbf{1}\left\{\text{ans}\left(\widehat{R}_i^{t+1}(a; \xi_i)\right) = y\right\}, \tag{6}$$

where the KEEP outcome is evaluated directly on R_i^t. We write $r_i(a)$ when the decoding seed is clear from context.

Receiver-conditioned intervention preferences are obtained from outcome differences within the matched checkpoint:

$$\mathcal{P}_i = \left\{(a^+, a^-) \in \mathcal{A}_i^t \times \mathcal{A}_i^t : r_i(a^+) > r_i(a^-)\right\}. \tag{7}$$

Only strict outcome differences produce directed preference pairs. Actions with the same observed outcome remain tied. Hence, if both SEND(j) and SELF yield a correct answer, neither action is preferred solely from that realization; similarly, two unsuccessful actions do not receive an artificial ordering.

The binary matched outcomes provide samples of the receiver utilities defined in Eq. (3). At the population level, these pairwise comparisons provide preference signals aligned with expected utility differences under the corresponding pairwise logistic objective. Individual matched comparisons can nevertheless be noisy when the action utilities are close.

4.2 Receiver-Conditioned Action Scoring

RPO-MAS assigns each candidate action a scalar score $f_\theta(z_i^t, a)$ using information available before the intervention is executed. The representation combines the question, the receiver's current

response, the candidate peer response for SEND actions, receiver–message interactions, and structural information from the public snapshot.

Let $e(\cdot) \in \mathbb{R}^{32}$ denote the textual representation obtained from a fixed MiniLM encoder followed by a PCA projection fitted on the training split. For each action a, define

$$e_a = \begin{cases} e(m_j^t), & a = \mathrm{M}_j, \\ \mathbf{0}, & a \in \{\mathrm{K}, \mathrm{S}\}. \end{cases}$$

The action representation is

$$\phi(z_i^t, a) = \Big[e(x); e(R_i^t); e_a; \mathbf{1}_{a \in \mathcal{M}_i^t}\big(e_a - e(R_i^t)\big); \mathbf{1}_{a \in \mathcal{M}_i^t}\big(e_a \odot e(R_i^t)\big); v(z_i^t, a) \Big], \tag{8}$$

where $\mathcal{M}_i^t = \mathcal{A}_i^t \setminus \{\mathrm{K}, \mathrm{S}\}$ and $v(z_i^t, a)$ contains pre-action structural information from $\mathcal{S}^t$, including the action type, answer-validity indicators, peer agreement, answer entropy, and consensus support. A shared MLP maps $\phi(z_i^t, a)$ to the scalar score $f_\theta(z_i^t, a)$. Ground-truth correctness and post-intervention outcomes are used only to construct supervision and are not included in the scoring features.

4.3 PREFERENCE OPTIMIZATION

The scorer is trained to assign larger values to preferred interventions. Preference pairs belong to three comparison families: *Update–Keep*, comparing an active intervention in $\{\mathrm{S}\} \cup \mathcal{M}_i^t$ with K; *Peer–Self*, comparing M_j with S; and *Peer–Peer*, comparing two peer-conditioned interventions.

The two preference orientations within the Update–Keep and Peer–Self families can occur at different frequencies. For family $k \in \{\text{Update–Keep}, \text{Peer–Self}\}$, let $M_{k,+}$ and $M_{k,-}$ denote the numbers of training pairs in the two orientations. For Update–Keep, the positive orientation denotes an active intervention preferred over KEEP; for Peer–Self, it denotes a peer intervention preferred over SELF. The reverse ordering defines the negative orientation. For a pair (a^+, a^-) belonging to family k with orientation $c(a^+, a^-) \in \{+, -\}$, we assign

$$w(a^+, a^-) = \left(\frac{M_{k,+} + M_{k,-}}{2M_{k,c(a^+,a^-)}} \right)^\alpha, \tag{9}$$

with $\alpha = 0.5$. Peer–Peer pairs receive unit weight. The orientation counts are computed from the training preference set and kept fixed during optimization.

The receiver-level preference loss is

$$\ell_{\mathrm{RPO}}(\theta; z_i^t) = \frac{1}{|\mathcal{P}_i|} \sum_{(a^+,a^-) \in \mathcal{P}_i} w(a^+, a^-) \log\left(1 + \exp\left(-\left[f_\theta(z_i^t, a^+) - f_\theta(z_i^t, a^-)\right]\right)\right), \tag{10}$$

with $\ell_{\mathrm{RPO}}(\theta; z_i^t) = 0$ when $\mathcal{P}_i = \varnothing$. The normalization is applied within each receiver checkpoint so that states inducing more strict pairs do not receive larger weight solely because they generate more comparisons.

We additionally regularize score magnitude with

$$\mathcal{R}(\theta; z_i^t) = \frac{1}{|\mathcal{A}_i^t|} \sum_{a \in \mathcal{A}_i^t} D_{\mathrm{KL}}\left(\mathrm{Bernoulli}\left(\sigma(f_\theta(z_i^t, a))\right) \,\|\, \mathrm{Bernoulli}(0.5)\right), \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function. Let $\mathcal{D}_{\mathrm{train}}$ denote the collection of training receiver checkpoints and their induced preference sets. The final objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{(z_i^t, \mathcal{P}_i) \sim \mathcal{D}_{\mathrm{train}}}\left[\ell_{\mathrm{RPO}}(\theta; z_i^t) + \lambda \mathcal{R}(\theta; z_i^t)\right], \tag{12}$$

with $\lambda = 0.01$.

4.4 SELECTIVE COMMUNICATION

At inference time, the scorer evaluates the available interventions from the current receiver context without access to ground-truth labels or post-intervention outcomes. For receiver i, we first identify the highest-scoring active intervention,

$$a_i^\star = \arg\max_{a \in \mathcal{A}_i^t \setminus \{\mathrm{K}\}} f_\theta(z_i^t, a). \tag{13}$$

The selected active intervention is executed only when its score exceeds that of state preservation by more than a threshold τ:

$$a_i^t = \begin{cases} a_i^\star, & f_\theta(z_i^t, a_i^\star) - f_\theta(z_i^t, \mathrm{K}) > \tau, \\ \mathrm{K}, & \text{otherwise.} \end{cases} \tag{14}$$

We use $\tau = 0$ as the canonical operating point and resolve exact threshold ties in favor of KEEP. When K is selected, receiver i retains R_i^t. Otherwise, only the selected SELF or SEND branch is executed to produce R_i^{t+1} according to Eq. (2).

All receivers make their decisions from the same public snapshot $\mathcal{S}^t$ and update synchronously to form $\mathcal{S}^{t+1}$. Interaction terminates early when every receiver selects KEEP. After the final round, the team prediction is obtained by majority vote over the agents' final parseable answers.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

(a) Datasets. We evaluate RPO-MAS on seven reasoning benchmarks: **Arithmetics**, **GSM8K** (Cobbe et al., 2021), **MMLU Professional Medicine**, **MMLU Formal Logic** (Hendrycks et al., 2021), **HellaSwag** (Zellers et al., 2019), **CommonsenseQA (CSQA)** (Talmor et al., 2019), and **HH-RLHF** (Bai et al., 2022). Preference data are constructed from training splits and evaluation is conducted on disjoint test splits. Dataset details are provided in Appendix A.

(b) Models and Baselines. We evaluate two instruction-tuned LLMs: Qwen2.5-7B-Instruct (Yang et al., 2024) and Llama-3.1-8B-Instruct (Dubey et al., 2024). Each system consists of five homogeneous agents sharing the same backbone. We compare against Single Agent, Majority Voting (Choi et al., 2025), Self-Refinement (Madaan et al., 2023), Decentralized MAD (Du et al., 2024), and Sparse MAD (Choi et al., 2025). Detailed baseline implementations follow Appendix B.

(c) Evaluation Protocol. All experiments use the same decoding configuration with temperature 1.0, top-$p = 0.9$, top-$k = 20$, repetition penalty 1.05, and maximum generation length of 512 tokens. RPO-MAS selects among KEEP, SELF, and SEND interventions for each receiver, with threshold $\tau = 0$. The system runs for up to two communication rounds and aggregates final answers by majority vote. Implementation details are provided in Appendix B.

(d) Evaluation Metrics. We report four metrics: **Team Acc.**, measuring final team prediction accuracy; **Rec. Acc.**, measuring post-intervention receiver accuracy; **Corr.**, measuring the fraction of incorrect receiver states corrected by executed interventions; and **Harm**, measuring the fraction of correct receiver states disrupted by executed interventions. Higher values indicate better performance for Team Acc., Rec. Acc., and Corr., while lower values indicate better performance for Harm. Detailed definitions are provided in Appendix C.

5.2 MAIN RESULTS

Table 1 presents the main results across seven reasoning benchmarks using Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct.

Overall Performance. RPO-MAS achieves the strongest average performance across both backbone models, obtaining 77.72% average accuracy with Qwen2.5-7B-Instruct and 73.13% with Llama-3.1-8B-Instruct. The improvements are consistent across both independent aggregation and communication-based baselines. Compared with Majority Voting, RPO-MAS improves average accuracy by 0.81 and 0.71 percentage points on Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, respectively, while achieving larger gains over Decentralized MAD (3.95 and 3.84 points) and Sparse MAD (4.42 and 3.23 points). These results indicate that the benefit of multi-agent communication does not come from increasing interaction alone, but from selecting receiver-specific interventions according to their expected consequences.

Comparison with Independent Aggregation. Majority Voting provides a strong non-communicative baseline by aggregating independent reasoning trajectories from multiple agents. RPO-MAS further improves upon this strategy while selectively incorporating peer information rather than requiring all agents to communicate. The consistent gains across GSM8K, MMLU, HellaSwag, and CSQA suggest

Table 1: Main performance comparison of RPO-MAS with existing multi-agent reasoning baselines across seven reasoning benchmarks. Accuracy (%) is reported for each task.

Backbone	Method	Arith.	GSM8K	Med.	Logic	HellaSwag	CSQA	HH-RLHF	Avg.
Qwen2.5 7B-Instruct	*No Communication*								
	Single Agent	81.40	87.13	78.68	49.05	78.80	81.53	47.73	72.05
	Majority Voting	99.00	94.00	79.41	53.97	80.33	83.00	48.67	76.91
	Self-Refinement	99.00	94.33	77.57	53.17	78.67	80.33	46.67	75.68
	Communication								
	Decentralized MAD ($T = 2$)	76.00	88.67	80.51	55.56	80.33	85.67	49.67	73.77
	Sparse MAD ($T = 2$)	84.00	90.33	80.51	47.62	79.67	83.67	47.33	73.30
	RPO-MAS	**99.00**	**95.00**	**80.51**	**55.56**	**81.00**	**84.67**	48.33	**77.72**
Llama-3.1 8B-Instruct	*No Communication*								
	Single Agent	73.20	73.93	74.41	37.94	62.67	67.67	44.40	62.03
	Majority Voting	89.00	87.33	81.99	51.59	69.33	78.00	49.67	72.42
	Self-Refinement	89.00	87.67	79.78	52.38	67.67	75.00	48.00	71.36
	Communication								
	Decentralized MAD ($T = 2$)	82.00	79.33	78.68	52.38	67.67	72.67	52.33	69.29
	Sparse MAD ($T = 2$)	85.00	81.33	80.15	46.83	67.67	75.67	52.67	69.90
	RPO-MAS	**89.00**	**88.33**	**82.72**	**53.17**	**70.00**	**79.33**	49.33	**73.13**

Table 2: Ablation study of RPO-MAS components on GSM8K and CSQA using Qwen2.5-7B-Instruct.

	Design			GSM8K				CSQA			
Variant	Rcv.	Outcome	Bal.	Team Acc.↑	Rec. Acc.↑	Corr. ↑	Harm ↓	Team Acc.↑	Rec. Acc.↑	Corr. ↑	Harm ↓
Source-based Scoring	–	✓	✓	93.33	91.20	45.67	12.32	82.67	80.87	11.93	3.17
Source-quality Supervision	✓	–	✓	93.67	91.80	41.26	8.96	83.00	81.60	10.40	2.34
w/o Directional Balancing	✓	✓	–	94.33	92.27	45.41	6.15	83.67	82.47	12.43	1.60
RPO-MAS	✓	✓	✓	**95.00**	**92.87**	**51.43**	**6.06**	**84.67**	**83.40**	**15.80**	**1.52**

that receiver-conditioned intervention selection can identify cases where peer information provides additional value beyond independent diversity. At the same time, the comparable performance on Arithmetics indicates that communication is not uniformly beneficial and that preserving existing reasoning states can be preferable when additional information provides limited value.

Comparison with Multi-Agent Debate. Compared with conventional debate-based communication, RPO-MAS obtains larger improvements, especially against methods that apply peer interaction more uniformly. For example, RPO-MAS improves average accuracy over Decentralized MAD by 3.95 and 3.84 points on the two backbones, respectively. This gap suggests that the main advantage of RPO-MAS comes from receiver-level selection rather than communication itself. By explicitly comparing peer-conditioned updating with autonomous refinement and state preservation, RPO-MAS avoids unnecessary updates while retaining communication opportunities when they are beneficial.

5.3 Ablation Study

We ablate three components of RPO-MAS on GSM8K and CSQA with Qwen2.5-7B-Instruct: receiver-conditioned scoring, receiver-side outcome supervision, and directional preference balancing. Table 2 reports both final accuracy and receiver-level correction and harm metrics.

Receiver conditioning reduces harmful updates. Replacing receiver-conditioned scoring with source-based scoring substantially increases harmful interventions. On GSM8K, Harm increases from 6.06% to 12.32%, while Rec. Acc. decreases from 92.87% to 91.20%. A similar degradation appears on CSQA. Since peer information remains available in this ablation, the performance drop indicates that source quality alone is insufficient for deciding whether a message is beneficial for a particular receiver state. This result supports the receiver-conditioned formulation of RPO-MAS: the same peer response may require different intervention choices depending on the current receiver state.

Table 3: Intervention selection behavior of RPO-MAS across seven reasoning benchmarks. The table reports the proportion of KEEP, SELF, and SEND decisions across interaction rounds for both backbone models.

	Qwen2.5-7B-Instruct						Llama-3.1-8B-Instruct					
	Round 1			Round 2			Round 1			Round 2		
Dataset	KEEP	SELF	SEND	KEEP	SELF	SEND	KEEP	SELF	SEND	KEEP	SELF	SEND
Arith.	93.20	0.60	6.20	98.00	0.20	1.80	88.40	0.60	11.00	95.20	0.20	4.60
GSM8K	84.73	0.73	14.54	92.00	0.27	7.73	78.93	0.80	20.27	86.93	0.33	12.74
Med.	59.41	1.40	39.19	68.31	0.81	30.88	54.85	1.32	43.83	63.60	0.74	35.66
Logic	50.95	3.33	45.72	61.59	2.06	36.35	47.30	2.86	49.84	57.30	1.75	40.95
HellaSwag	47.87	0.93	51.20	56.20	0.53	43.27	43.47	0.87	55.66	51.73	0.47	47.80
CSQA	37.13	1.20	61.67	43.00	0.73	56.27	33.20	1.13	65.67	38.93	0.67	60.40
HH-RLHF	68.00	0.53	31.47	78.33	0.33	21.34	64.53	0.60	34.87	73.87	0.40	25.73

Receiver-side outcomes improve corrective decisions. Replacing outcome-based supervision with source-quality supervision mainly reduces the ability to identify useful interventions. The correction rate decreases from 51.43% to 41.26% on GSM8K, with consistent degradation on CSQA. This result highlights that receiver-side outcomes provide a more suitable basis for learning intervention preferences than source-level message properties.

Directional balancing improves the correction–protection trade-off. Removing directional balancing leads to lower correction rates while only slightly changing Harm. For example, Corr. decreases from 51.43% to 45.41% on GSM8K, whereas Harm remains nearly unchanged. This suggests that directional balancing improves preference learning under asymmetric comparison distributions, helping the scorer better capture the relative preference between intervention directions without making the policy uniformly conservative.

5.4 COMMUNICATION BEHAVIOR

To characterize how RPO-MAS performs selective communication, we analyze the intervention decisions across tasks and interaction rounds. Table 3 reports the proportions of KEEP, SELF, and SEND under the two backbone models.

RPO-MAS learns distinct intervention patterns across tasks. Arithmetics is dominated by KEEP, while CSQA, HellaSwag, and MMLU Formal Logic invoke SEND more frequently. This difference suggests that the learned policy does not apply communication uniformly, but adapts intervention choices according to the receiver states and reasoning requirements induced by different tasks.

The intervention distribution also changes as interaction progresses. Across tasks and both backbone models, KEEP becomes more prevalent in the later round, while SEND consistently decreases. This pattern indicates that RPO-MAS reassesses the value of additional interventions after receiver states evolve: communication is favored when it can provide useful correction, while preservation becomes preferable once further updates are less likely to improve the current state. The small proportion of SELF further suggests that, when an active update is selected, peer-conditioned reasoning is more frequently preferred than additional autonomous refinement.

6 CONCLUSION

In this paper, we introduced RPO-MAS, a receiver-conditioned framework for selective multi-agent communication that learns intervention preferences from receiver-side outcomes. By comparing peer-conditioned updating, autonomous refinement, and state preservation from the same receiver state, RPO-MAS selects interventions according to their observed consequences. Experiments across seven reasoning benchmarks using two LLM backbones show that RPO-MAS improves team performance over independent aggregation and communication-based baselines. Our analyses further show that receiver conditioning and execution-based supervision play important roles in improving correction while reducing harmful changes to correct receiver states, and that directional preference balancing improves the correction–protection trade-off. The learned policy also adapts intervention choices across tasks and interaction rounds, reducing unnecessary updates as receiver states evolve. These results provide evidence that receiver-side outcomes offer an effective basis for selective multi-agent communication.

AI USE STATEMENT

We use generative artificial intelligence for article language translation and polishing, as well as to assist in researching some related background articles.

REPRODUCIBILITY STATEMENT

Section 4 specifies the intervention space, receiver-side supervision, scoring model, training objective, and inference procedure. Appendix A documents the datasets and evaluation splits, Appendix B provides the controller training, inference protocol, generation settings, and prompt templates, and Appendix C gives the complete definitions of the evaluation metrics. We additionally report the model backbones, decoding parameters, training and validation sizes, and intervention-selection settings used in our experiments.

ETHICS STATEMENT

The experiments use public reasoning benchmarks and do not involve human subjects. A controller that learns when to accept a peer response can still propagate incorrect or biased information; benchmark correctness does not establish reliability in consequential applications. Dataset and model licenses will accompany any released artifacts. Historical benchmark use and the provenance of externally published results are disclosed in the experimental appendix.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nitya DasSarma, Dawn Drain, Deep Ganguli, Tom Henighan, Nicholas Joseph, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.601. URL https://aclanthology.org/2025.findings-acl.601/.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499. PMLR, 2025b. URL https://proceedings.mlr.press/v267/chen25bx.html.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3405. URL https://papers.nips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-MAD: Consensus-free multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31977–31997. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1600. URL https://aclanthology.org/2026.findings-acl.1600/.

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, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL https://arxiv.org/abs/2305.14325.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Akhil Letman, Akhil Mathur, Alexander Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wei Fan, JinYi Yoon, and Bo Ji. iMAD: Intelligent multi-agent debate for efficient and accurate LLM inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29403–29411, 2026. doi: 10.1609/aaai.v40i35.40181. URL [https://ojs.aaai.org/index.php/AAAI/article/view/40181.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL [https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.606. URL [https://aclanthology.org/2025.findings-acl.606/.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, JD Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025. URL [https://proceedings.iclr.cc/paper_files/paper/2025/hash/871ac99fdc5282d0301934d23945ebaa-Abstract-Conference.html.

Zijie Lin and Bryan Hooi. Enhancing multi-agent debate system performance via confidence expression. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6453–6471. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.343. URL [https://aclanthology.org/2025.findings-emnlp.343/.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *International Conference on Learning Representations*, 2025. URL [https://proceedings.iclr.cc/paper_files/paper/2025/hash/3de667dab3b3d812583abc0a786139a0-Abstract-Conference.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Rafael Rafailov, Ivan Laptev, Philip Torr, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. MALT: Improving reasoning with multi-agent LLM training. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=jXP9bgFack.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. URL https://aclanthology.org/N19-1421/.

Hongduan Tian, Xiao Feng, Ziyuan Zhao, Xiangyu Zhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/09425891e393e64b0535194a81ba15b7-Abstract-Conference.html.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y. Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, 2025. URL [https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434be94e82c54327bb9dcaf7fca52b6-Abstract-Conference.html.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *International Conference on Learning Representations*, 2025. URL [https://proceedings.iclr.cc/paper_files/paper/2025/hash/1da9ca7e9cef4b1af63913f05d1630a4-Abstract-Conference.html.

Hao Wu, Shoucheng Song, Chang Yao, Haoyu Wang, Huaiyu Wan, Youfang Lin, and Kai Lv. Group perspective matters: Regulating debate relationships can mitigate blind conformity in multi-agent debate. *arXiv preprint arXiv:2608.03648*, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/P19-1472/.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S^2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9393–9408, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.475. URL https://aclanthology.org/2025.naacl-long.475/.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

Ruiwen Zhou, Maojia Song, Xiaobao Wu, Sitao Cheng, Xunjian Yin, Yuxi Xie, Zhuoqun Hao, Wenyue Hua, Liangming Pan, Soujanya Poria, and Min-Yen Kan. Epistemic context learning: Building trust the right way in LLM-based multi-agent systems, 2026. URL https://arxiv.org/abs/2601.21742.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.195. URL [https://aclanthology.org/2025.acl-long.195/.

Table 4: Evaluation datasets and splits used in our experiments.

Dataset	Evaluation Split	# Instances
Arithmetics	Test	100
GSM8K	Test	300
MMLU Medicine	Test	272
MMLU Logic	Test	126
HellaSwag	Test	300
CSQA	Validation	300
HH-RLHF	Test	300

A Dataset Details

We follow the benchmark suite and evaluation subsets of Choi et al. (2025), covering seven reasoning and decision-making benchmarks: Arithmetics, GSM8K (Cobbe et al., 2021), MMLU Professional Medicine, MMLU Formal Logic (Hendrycks et al., 2021), HellaSwag, CommonsenseQA (CSQA) (Talmor et al., 2019), and HH-RLHF.

For each dataset and backbone, we independently train a task-specific RPO-MAS controller. We use 400 training instances to construct receiver-side preference data and 100 disjoint validation instances for model selection. Evaluation follows the corresponding benchmark subset used in Choi et al. (2025). Table 4 summarizes the evaluation splits and numbers of instances.

For multiple-choice benchmarks, correctness is determined by exact matching between the parsed option and the reference label. For GSM8K and Arithmetics, the final numerical answer is extracted from the model response and compared with the reference answer. For HH-RLHF, the parsed preference label is compared with the reference preference. The same answer parsing procedure is used across all methods evaluated on the same dataset.

B Experimental Details

B.1 Controller Training

For each dataset and backbone, we independently train a task-specific RPO-MAS controller. No controller parameters are shared across datasets or backbone models. The underlying language models remain frozen throughout controller training.

For each training question, five agents first produce receiver states and candidate peer responses using the corresponding backbone model. At each receiver checkpoint, we evaluate the complete intervention set

$$\mathcal{A}_i^t = \{\textsc{Keep}, \textsc{Self}\} \cup \{\textsc{Send}(j) : j \neq i\}.$$

All candidate interventions are evaluated from the same pre-action receiver state. KEEP directly preserves the current response, SELF performs an additional autonomous reasoning step, and SEND performs an additional reasoning step conditioned on the selected peer response.

For each action $a \in \mathcal{A}_i^t$, its realized receiver-side outcome is determined by the correctness of the resulting receiver answer. Strict outcome differences generate directed preference pairs, whereas actions with equal outcomes remain unordered. These preferences are used to optimize the receiver-conditioned scorer $f_\theta(z_i^t, a)$ following the objective in Section 4.

The textual components of the controller input are encoded using a frozen MiniLM encoder. The resulting representations are projected to 32 dimensions using PCA fitted only on the corresponding training split. For each candidate action, the controller combines the question representation, current receiver representation, candidate peer representation, receiver–peer interaction features, and the pre-action structural features defined in Section 4. A shared two-hidden-layer MLP maps the resulting representation to a scalar action score.

Directional preference balancing is applied to the Update–Keep and Peer–Self comparison families. We use the directional weighting exponent $\alpha = 0.5$ and score-regularization coefficient $\lambda = 0.01$,

Table 5: Principal implementation settings for RPO-MAS.

Setting	Value
Number of agents	5
Backbones	Qwen2.5-7B-Instruct; Llama-3.1-8B-Instruct
Controller scope	Dataset- and backbone-specific
Training instances per controller	400
Validation instances per controller	100
Text encoder	MiniLM
Text encoder update	Frozen
PCA dimension	32
Directional exponent α	0.5
Regularization coefficient λ	0.01
Maximum interaction rounds	2
Intervention threshold τ	0
Threshold tie handling	KEEP
Temperature	1.0
Top-p	0.9
Top-k	20
Repetition penalty	1.05
Maximum new tokens	512
Final aggregation	Majority vote

following the definitions in Section 4. Controller selection is performed using receiver accuracy on the corresponding validation split.

B.2 MULTI-AGENT INFERENCE PROTOCOL

All experiments use $N = 5$ homogeneous agents sharing the same backbone model. Each question is first answered independently by the five agents, forming the initial public snapshot S^0.

At interaction round t, the task-specific controller evaluates all available interventions for each receiver i from its current context z_i^t. With five agents, each receiver has six candidate actions: one KEEP, one SELF, and four SEND(j) actions corresponding to the other four agents.

For each receiver, the controller first identifies the highest-scoring active intervention

$$a_i^\star = \arg \max_{a \in \mathcal{A}_i^t \setminus \{\text{KEEP}\}} f_\theta(z_i^t, a).$$

The active intervention is executed when

$$f_\theta(z_i^t, a_i^\star) - f_\theta(z_i^t, \text{KEEP}) > \tau.$$

We use $\tau = 0$ throughout the experiments and resolve exact ties in favor of KEEP.

All receivers make their decisions from the same public snapshot and update synchronously. RPO-MAS runs for at most two adaptive interaction rounds and terminates early when all receivers select KEEP. After the final interaction round, the team prediction is obtained by majority vote over the five agents' final parseable answers.

B.3 GENERATION AND IMPLEMENTATION SETTINGS

We use Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct as the backbone language models. All agents within one multi-agent system use the same backbone and decoding configuration.

For generation, we use temperature 1.0, top-$p = 0.9$, top-$k = 20$, a repetition penalty of 1.05, and a maximum of 512 newly generated tokens. The same generation configuration is used for the initial responses and all active interventions.

The principal experimental settings are summarized in Table 5.

B.4 PROMPT TEMPLATES

All agents sharing a backbone use the same task-specific base prompt. The intervention templates differ only in the information available to the receiver. SELF exposes the receiver to its own current response and requests an additional independent reasoning step, whereas SEND additionally provides the selected peer response. KEEP performs no additional generation.

For the initial response, the model receives the task question and is instructed to solve the problem and provide a parseable final answer. For SELF, the model receives the original question together with its current response and is instructed to reconsider the problem independently before producing an updated answer. For SEND(j), the model additionally receives agent j's current response and is instructed to reconsider its answer using the peer response as additional information.

The same intervention templates are used for controller-data construction and test-time execution.

C EVALUATION METRICS

We evaluate RPO-MAS at both the team and receiver levels. Let $\mathcal{Q}$ denote the evaluation set, N the number of receivers, and T_q the number of executed interaction rounds for question q. Let $c_{q,i}^{t,-} \in \{0,1\}$ and $c_{q,i}^{t,+} \in \{0,1\}$ denote the correctness of receiver i immediately before and after its action at round t, respectively.

Team Accuracy. Team Accuracy measures the correctness of the final team prediction after multi-agent aggregation:

$$\mathrm{TeamAcc} = \frac{\sum_{q\in\mathcal{Q}} \mathbf{1}\left[\hat{y}_q^{\mathrm{team}} = y_q\right]}{|\mathcal{Q}|}. \tag{15}$$

Here, $\hat{y}_q^{\mathrm{team}}$ is obtained by majority vote over the agents' final parseable answers.

Receiver Accuracy. Receiver Accuracy measures the average correctness of post-action receiver states across all receivers and executed interaction rounds:

$$\mathrm{RecAcc} = \frac{\sum_{q\in\mathcal{Q}} \sum_{t=1}^{T_q} \sum_{i=1}^{N} c_{q,i}^{t,+}}{\sum_{q\in\mathcal{Q}} T_q N}. \tag{16}$$

The metric therefore evaluates the receiver-level state quality induced by the controller decisions throughout the interaction process.

Correction Rate. Correction Rate measures the effectiveness of actually executed interventions in correcting erroneous receiver states. Define the set of active interventions as

$$\mathcal{I} = \left\{(q,i,t) : a_{q,i}^t \neq \textsc{Keep}\right\}. \tag{17}$$

Then,

$$\mathrm{Corr} = \frac{\sum_{(q,i,t)\in\mathcal{I}} \mathbf{1}\left[c_{q,i}^{t,-} = 0 \wedge c_{q,i}^{t,+} = 1\right]}{\sum_{(q,i,t)\in\mathcal{I}} \mathbf{1}\left[c_{q,i}^{t,-} = 0\right]}. \tag{18}$$

Thus, the denominator contains only executed SELF and SEND interventions whose receiver state was incorrect before intervention.

Harm Rate. Harm Rate measures how often an actually executed intervention changes a correct receiver state into an incorrect one:

$$\mathrm{Harm} = \frac{\sum_{(q,i,t)\in\mathcal{I}} \mathbf{1}\left[c_{q,i}^{t,-} = 1 \wedge c_{q,i}^{t,+} = 0\right]}{\sum_{(q,i,t)\in\mathcal{I}} \mathbf{1}\left[c_{q,i}^{t,-} = 1\right]}. \tag{19}$$

The denominator therefore contains only executed SELF and SEND interventions whose receiver state was correct before intervention.

Team Acc., Rec. Acc., and Corr. are better when higher, whereas Harm is better when lower.