MARO: Learning Stronger Reasoning from Social Interaction

Anonymous ACL submission

Abstract

Humans face countless scenarios that require reasoning and judgment in daily life. However, existing large language model training methods primarily allow models to learn from existing textual content or solve predetermined problems, lacking experience in real scenarios involving interaction, negotiation, and competition with others. To address this, this paper proposes Multi-Agent Reward Optimization (MARO), a method that enables large language models (LLMs) to acquire stronger reasoning abilities by learning and practicing in multi-agent social environments. Specifically, MARO first addresses the sparse learning signal problem by decomposing final success or failure outcomes into each specific behavior during the interaction process; second, it handles the uneven role distribution problem by balancing the training sample weights of different roles; finally, it addresses environmental instability issues by directly evaluating the utility of each behavior. Experimental results demonstrate that MARO not only achieves significant improvements in social reasoning capabilities, but also that the abilities acquired through social simulation learning can effectively transfer to other tasks such as mathematical reasoning and instruction following. This reveals the tremendous potential of multi-agent social learning in enhancing the general reasoning capabilities of LLMs.

1 Introduction

During daily life, humans face countless scenarios that require reasoning and judgment: evaluating complex information to make decisions, seeking balance among multiple interests, understanding others' intentions and formulating corresponding strategies. These activities all require advanced cognitive abilities—observing the environment, analyzing information, predicting others' behavior, developing strategies, and adjusting decisions based on feedback. It is through repeated practice and learning in these real social scenarios that humans gradually develop exceptional problem-solving capabilities.

The ability to reason through various complex problems is a core skill that humans have honed through long-term social practice, and it is also an important goal pursued by Large Language Models (LLMs). Through large-scale pre-training and fine-tuning on various tasks, LLMs have demonstrated exceptional performance in natural language understanding, knowledge question-answering, and text generation (Yin et al., 2024; Qu et al., 2025; Cheng et al., 2024; Zhang et al., 2024; Shinn et al., 2023; Guo et al., 2024). However, current training approaches mainly allow LLMs to learn from existing textual content or solve predetermined problems, lacking experience in real-life scenarios involving interaction, negotiation, and competition with others. Compared to symbolic reasoning in domains like mathematics and coding, life-oriented reasoning involves broader knowledge integration, situational understanding, and strategic planning. This gap in training paradigms limits the further generalization of LLM reasoning capabilities (Wang et al., 2025a).

To enable LLMs to acquire stronger capabilities, a natural approach is to let them learn and practice in real social scenarios, just like humans do (Hong et al., 2024; Cheruiyot et al., 2025). In such learning environments, multiple agents exist simultaneously, each with their own tasks to complete or interests to protect. LLMs need to play the role of one of these agents, communicating, negotiating, and even competing with other agents who also have independent goals and strategies (Li et al., 2025; Liu et al., 2024b; Patil et al., 2023). LLMs must maximize their goal achievement by observing others' words and actions, understanding their intentions, and formulating their own action plans. This learning approach requires LLMs not

*Corresponding Author

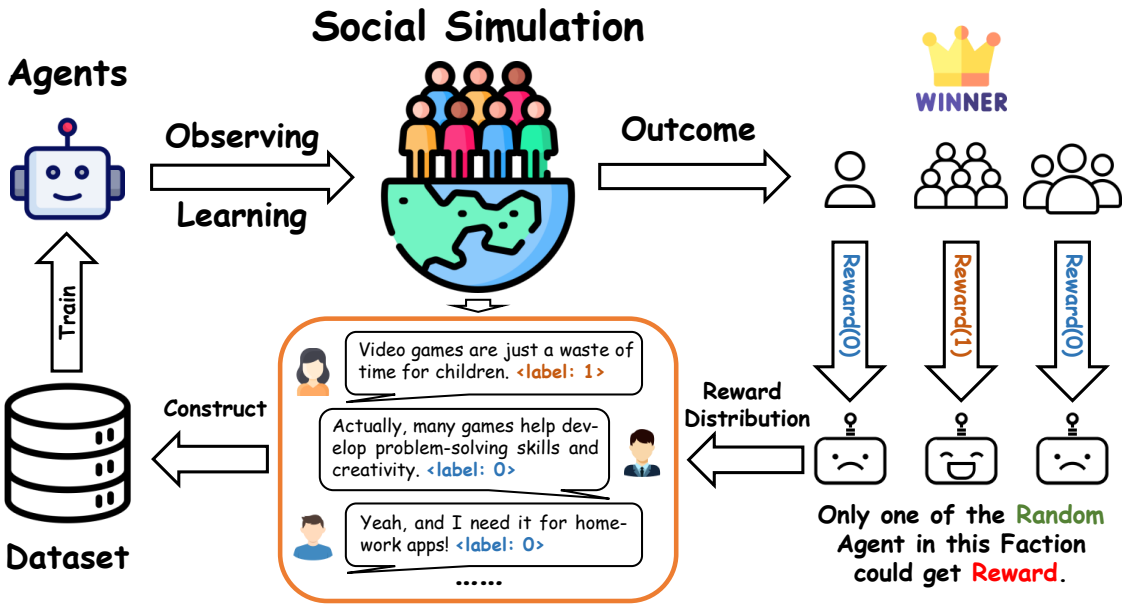


Figure 1: Overview of the MARO workflow. First, agents interact in social scenarios where each agent pursues individual goals through communication and decision-making. Second, upon interaction completion, the system evaluates final outcomes to determine success or failure for each participating agent. Third, MARO decomposes these sparse final outcomes into dense step-wise rewards distributed across each agent's action trajectory, incorporating role-specific weights for balanced training. Fourth, the decomposed reward data is used to optimize LLMs through MARO's specialized loss function for enhanced social reasoning capabilities.

only to understand language but also to learn how to make wise decisions in dynamically changing environments, handle conflicts of interest, and even change strategies when necessary to adapt to new situations (Jin et al., 2025; Maldonado et al., 2024).

However, training LLMs in multi-agent environments faces three key challenges (Liu et al., 2024a). First is the sparse learning signal problem: for individual LLMs, during long interaction processes, only the final success or failure results are clear, but LLMs need to know which specific decisions during the lengthy interaction process were effective and which were harmful (Cai et al., 2025a). Second is the uneven role distribution problem: in many social scenarios, the difficulty of success varies greatly among different roles. For example, in business negotiations, sellers often find it easier to close deals than buyers, and in team collaboration, leaders are more likely to achieve good results than ordinary members. This leads to certain roles generating numerous success cases while other roles have scarce successful experiences, causing training data to be heavily biased toward roles that are naturally easier to succeed in (ShengbinYue et al., 2025; Karten et al., 2025). Third is the environment instability problem: in real social scenarios, the circumstances of each interaction are never completely identical—the number of agents, their personalities, and goals may all change. While LLM training typically relies on relatively stable feedback patterns, this constantly changing environment leads to inconsistent gradient directions during training, making it difficult for models to converge to optimal solutions (Zhang, 2025).

To address these challenges, we propose Multi-Agent Reward Optimization (MARO). MARO decomposes the final success/failure results into each specific behavior throughout the interaction process to tackle the sparse learning signal problem, enabling LLMs to clearly identify the value of each decision step. MARO balances the training sample weights of different roles to address the uneven role distribution problem, ensuring that LLMs can learn optimal strategies for various roles rather than just the behavioral patterns of naturally successful roles. MARO adopts a more robust training approach that directly evaluates the utility of each behavior to handle the environment instability problem, rather than performing complex strategy comparisons, reducing the interference of environmental changes on training consistency.

To validate the effectiveness of MARO, we con-

Symbol	Meaning
G	Multi-agent interaction instance $(E, \mathcal{A}, S, O)$
$s_t^{a_i}, o_t^{a_i}$	State and action of agent a_i at time t
$Success(a_i, G)$	Binary indicator of goal achievement for agent a_i
C_{win}, C_{lose}	Sets of agents in winning and losing camps
$w(camp(a_i))$	Camp-specific balancing weight for loss calculation
$\pi_\theta, \pi_{\text{ref}}$	Learned policy and reference policy
$r_\theta(o_t^{a_i})$	Implicit reward derived from policy π_θ
z_0	Adaptive KL divergence reference point
β	Hyperparameter controlling risk aversion

Table 1: Key symbols in MARO

ducted extensive experiments in simulated social environments. The results demonstrate that MARO significantly enhances LLMs' reasoning abilities across multiple dimensions, with improvements in social reasoning, transferable gains to mathematical reasoning and instruction-following tasks, and superior performance compared to traditional supervised fine-tuning methods. Importantly, by learning in social environments, LLMs can acquire transferable general reasoning abilities; compared with relying on high-quality, human-curated data, large-scale synthetic data from simulated environments offers a more favorable performance–cost trade-off. Moreover, complex social environments are more effective than simple ones at fostering improvements in general reasoning.

2 Related Work

2.1 Traditional Multi-Agent Systems and Reinforcement Learning

Traditional Multi-Agent Systems originate from game theory and distributed artificial intelligence (Shapley, 1953; Littman, 1994; Busoniu et al., 2008). Multi-agent reinforcement learning (MARL) introduces mathematical models like Markov games and Dec-POMDP (Oliehoek et al., 2016; Busoniu et al., 2008). Representative algorithms include Independent Q-Learning (IQL) (Tan, 1993), Value-Decomposition Networks (VDN) (Sunehag et al., 2017), QMIX (Rashid et al., 2018), MADDPG (Lowe et al., 2017), and COMA (Foerster et al., 2018). The centralized training with decentralized execution (CTDE) paradigm (Zhang et al., 2021) has enabled practical applications. Recent surveys (Jin et al., 2025) comprehensively analyze multi-agent cooperative decision-making approaches. However, traditional MARL algorithms face severe limitations in reward signal generation, typically relying on sparse, manually designed reward functions that fail to capture multi-agent interaction complexity (Hernandez-Leal et al., 2019; Rashid et al., 2018). Our MARO method addresses reward sparsity through trajectory expansion that automatically redistributes environmental rewards, enabling dense feedback generation.

2.2 LLM-Driven Multi-Agent Systems and Simulation-Based Learning

Large Language Models have demonstrated remarkable potential in environmental perception and reasoning-based decision-making (Achiam et al., 2023; Touvron et al., 2023; Xi et al., 2023; Guo et al., 2024; Gu et al., 2024b), advancing role-playing capabilities (Chen et al., 2024; Gu et al., 2024a). LLMs show human-like behaviors in social simulation (Park et al., 2023; Gu et al., 2024a; Wang et al., 2024a), policy simulation (Xiao et al., 2023), and game simulation (Xu et al., 2023b). Notable frameworks include AutoGen (Wu et al., 2024a), CAMEL (Li et al., 2023a), Stanford's Generative Agents (Park et al., 2023), MetaAgents (Li et al., 2023c), and AgentSims (Lin et al., 2023). Recent work explores theory-of-mind in multi-agent cooperation (Li et al., 2023b) and actor-critic frameworks for multi-agent reasoning (Wang et al., 2025b). Evaluation platforms like MIRAGE (Cai et al., 2025b), AgentSociety (Piao et al., 2025), and Sotopia (Zhou et al., 2023b) assess LLM agents in complex scenarios. Board games, particularly murder mysteries (Wu et al., 2023), Werewolf (Xu et al., 2023b,a; Shibata et al., 2023; Wu et al., 2024b), and Avalon (Wang et al., 2023), provide structured evaluation environments (Ye et al., 2025). However, these systems face critical challenges in reward distribution fairness across different agent factions, where dominant factions generate more positive samples while disadvantaged factions suffer from negative sample scarcity (Zhou et al., 2023b; Cai et al., 2025b; Park et al., 2023). Our MARO method introduces multi-faction reward balancing to ensure balanced positive and negative samples.

2.3 Reward and Preference Learning for LLMs

Alignment techniques are crucial for LLM development. RLHF (Ouyang et al., 2022; Christiano et al., 2017) and variants like RLAIF (Lee et al.,

Dataset	Method	Interaction		Persona		Trust		Investigation		Victory	
		killer ↑	Villager ↑	killer ↑	Villager ↑	killer ↑	Villager ↑	killer ↓	Villager ↑	killer ↑	Villager ↑
Simple (SOO)	Vanilla	57.88	50.21	42.51	41.83	38.60	35.83	1.77	3.09	59.31	40.69
	SFT	54.06	46.96	42.57	39.69	37.77	35.28	0.74	3.93	60.00	40.00
	-killer	49.38	51.08	38.86	41.18	35.82	36.01	**0.00**	4.10	60.00	40.00
	-Villager	53.88	48.23	40.59	40.88	35.63	34.41	2.41	4.52	60.37	39.63
	MAKTO	64.38	58.55	**49.39**	44.94	38.18	38.26	2.47	13.38	66.61	33.39
	-killer	64.56	52.78	48.04	44.15	36.87	37.61	1.00	7.96	71.92	28.08
	-Villager	61.50	58.36	44.05	45.56	37.92	40.02	2.25	14.80	71.06	28.94
	MARO (Ours)	**66.06**	60.05	48.40	46.61	41.90	**42.20**	1.53	16.36	70.65	29.35
	-killer	58.25	49.37	45.86	41.87	**43.40**	37.80	1.53	7.82	58.86	**41.14**
	-Villager	59.25	**60.53**	45.23	**46.69**	39.71	41.76	0.37	**18.21**	**74.40**	25.60
Complex (MUC)	Vanilla	55.25	53.49	39.60	39.78	38.29	34.13	0.97	15.87	41.92	**58.08**
	SFT	47.50	52.29	40.02	38.69	32.35	31.80	0.72	4.84	57.42	42.58
	-killer	50.67	53.33	37.63	38.76	34.29	33.93	**0.33**	11.29	53.67	46.33
	-Villager	51.08	50.32	36.70	39.28	37.14	33.94	2.76	3.72	53.70	46.30
	MAKTO	52.58	54.95	41.22	39.38	33.09	36.74	1.60	13.47	57.82	42.18
	-killer	55.50	54.11	40.47	40.59	35.59	34.31	4.59	14.44	57.74	42.26
	-Villager	54.75	53.60	38.70	40.65	39.56	35.56	3.36	8.15	49.54	50.46
	MARO (Ours)	55.08	**57.48**	43.38	41.81	44.56	42.84	2.58	12.84	**59.20**	40.80
	-killer	56.17	53.90	40.61	39.89	**44.65**	33.64	2.08	13.55	53.12	46.88
	-Villager	**57.83**	56.72	**43.65**	**43.71**	37.97	**44.27**	4.63	**17.86**	57.32	42.68

Table 2: Social Capability evaluation on Simple (SOO) and Complex (MUC) datasets. The table reports scores (in %) across five metrics, calculated separately for the Killer and Villager roles. The symbols ↑ and ↓ indicate that higher and lower scores are better, respectively. **Bold** and underlined values denote the best and second-best results. Regarding the setup: Base methods (e.g., **MARO**) indicate that **both roles are controlled by the same method. Rows with suffixes (e.g., *-killer* or *-Villager*) indicate that only the specified role is controlled by the method**, while the opponent is played by **Vanilla**.

2023) use reward models with reinforcement learning (e.g., PPO (Schulman et al., 2017)). Direct Preference Optimization (DPO) (Rafailov et al., 2023) transforms RL training into supervised learning, while KTO (Ethayarajh et al., 2024) and MRPO (Le et al., 2025) enhance generalization. Recent surveys (Du et al., 2025) analyze optimization techniques for LLM agents. However, these methods face a fundamental bottleneck: the prohibitively high cost of generating quality pairwise preference data in multi-agent scenarios (Dubois et al., 2023). The annotation burden becomes particularly severe when human evaluators must assess complex multi-turn interactions between multiple agents, requiring deep domain expertise and continuously substantial time investment. Our MARO method overcomes this through reward state optimization that trains models using binary labels on single samples via log-likelihood optimization, avoiding costly pairwise preference data while maintaining effective learning performance.

3 Multi-Agent Reward Optimization

As illustrated in Figure 1, MARO operates through a systematic process for training LLMs in multi-agent environments. MARO begins with multi-agent interactions where agents engage in social scenarios such as negotiations or collaborations, with each agent pursuing individual objectives while their actions and states are continuously recorded. After interaction completion, the system evaluates final outcomes to determine which agents successfully achieved their goals, providing binary success/failure signals. The core component involves decomposing these sparse final outcomes into dense, step-wise rewards that are distributed back across each agent's entire action sequence, with role-specific balancing applied to create a comprehensive training dataset. Finally, LLMs are optimized using this decomposed reward data through MARO's specialized loss function, enabling the models to learn effective social reasoning strategies from the rich multi-agent interaction experiences. The specific implementation methods are as follows.

Let a multi-agent interaction instance be denoted as $G = (E, \mathcal{A}, S, O)$, where E represents the environment state, $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ represents the set of all agents, $S = \{s_t^{a_i}\}$ represents the sequence of states observed by agent a_i at time step t (where $t \in \mathcal{T}$ and $\mathcal{T}$ is the time index set), and $O = \{o_t^{a_i}\}$ represents the sequence of actions taken

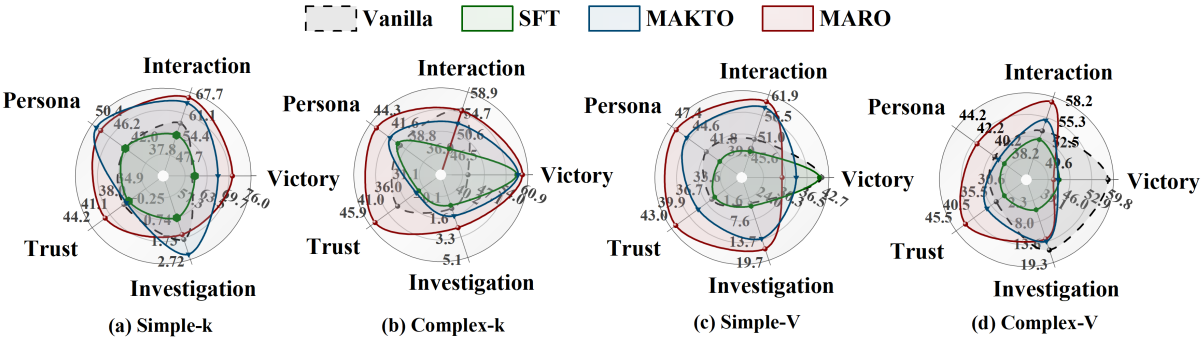


Figure 2: Performance comparison across different model types in various scenarios. The radar charts show five evaluation metrics (Persona, Interaction, Victory, Investigation, Trust) across four conditions: (a) **Simple-k**: killer faction performance in simple scenarios, (b) **Complex-k**: killer faction performance in complex scenarios, (c) **Simple-V**: Victim faction performance in simple scenarios, and (d) **Complex-V**: Victim faction performance in complex scenarios. All models (**Vanilla**, **SFT**, **MAKTO**, **MARO**) play both killer and Victim roles in their respective configurations.

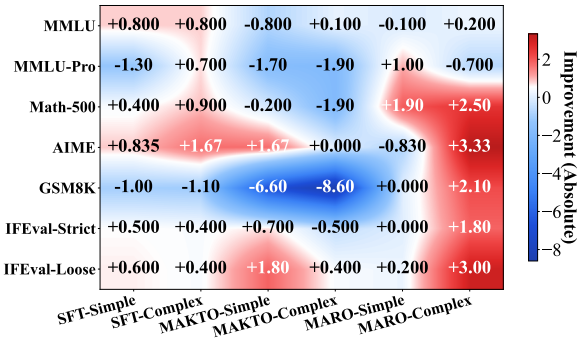


Figure 3: Comprehensive performance improvement heatmap showing absolute percentage point changes compared to Vanilla across all benchmarks.

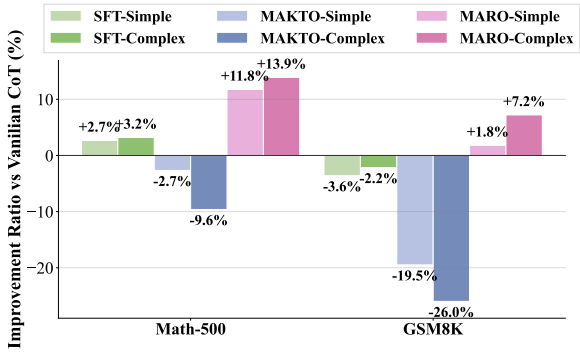


Figure 4: Relative improvement ratios of different training methods compared to Vanilla on mathematical reasoning tasks.

by agent a_i at time step t.

Each agent $a_i \in \mathcal{A}$ has an individual objective g_i, and the overall interaction outcome is determined by the collective actions of all agents. The success of agent a_i in achieving its objective is denoted by $Success(a_i, G)$:

$$Success(a_i, G) = \begin{cases} 1, & \text{if } a_i \text{ achieves } g_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Let C_{win} and C_{lose} represent the sets of agents in the winning and losing camps respectively, and $|C_{win}|$ and $|C_{lose}|$ represent the numbers of agents in the winning and losing camps respectively. To ensure balance, the number of agents eligible for rewards in each camp is:

$$N_{reward} = \min(|C_{win}|, |C_{lose}|) \quad (2)$$

For camps with more than N_{reward} agents, N_{reward} agents are randomly selected for reward allocation. Let $\mathcal{A}_{selected}$ denote the set of agents selected after balance filtering.

For each selected agent $a_i \in \mathcal{A}_{selected}$ and its action sequence $\{o_t^{a_i}\}$, we directly assign the final success signal to each action to obtain the step-wise reward:

$$\tilde{r}_t^{a_i}(o_t^{a_i}) = Success(a_i, G) \quad (3)$$

This avoids computing partial derivatives on discrete action sequences, which do not have well-defined gradients.

The training dataset is constructed as $\mathcal{D} = \{(o_t^{a_i}, y_t^{a_i})\}$, where the step-wise reward label is directly

$$y_t^{a_i} = \tilde{r}_t^{a_i}(o_t^{a_i}) = Success(a_i, G). \quad (4)$$

Let π_θ denote the language model policy and π_{ref} denote the reference policy. For any action $o_t^{a_i}$,

Script: Raging Sea of Rest Life (Role: Annie)		Script: Night at the Museum (Role: Curio He)	
Vanilla	**MARO**	**Vanilla**	**MARO**
[Self-introduction] **THOUGHT:** I need to display Annie's character traits... but I cannot directly expose my connection to Old Black's death. **RESPONSE:** Hello everyone, my name is Annie... My parents used to work at Mr. Cavill Lauren's company... Especially my mother, due to the doctor's negligence, she did not receive timely treatment and eventually left us. This incident still haunts me... *(Analysis: Mechanical introduction; superficial grasp of clues.)*	**[Self-introduction]** **THOUGHT:** Annie's character background is full of tragedy and revenge... I need to maintain a certain sense of mystery... **RESPONSE:** Hello everyone, I am Annie... Mr. Cavill, your management of Lauren Shipyard was commendable, but sadly, it eventually met its demise. At that time, my entire family were victims... I believe everyone has an unforgettable scar in their hearts, and I am no exception. Perhaps because of these experiences, I've learned to cherish the peaceful life I have now... *(Analysis: Uses emotive metaphors; shrewdly controls information.)*	**[Inquiry to Uncle Bai]** **THOUGHT:** I need to maintain mysteriousness... Asking Uncle Bai might provide more information... **RESPONSE:** [Inquiry] [Uncle Bai]: Mr. Bai, you are well-versed in Genghis Khan's artifacts. What secrets do you think these artifacts might hold? *(Analysis: Direct, generic question without strategic value.)*	**[Inquiry to Uncle Bai]** **THOUGHT:** ...I need to collect more evidence to prove that my father's death may be related to this... **RESPONSE:** [Inquiry] [Uncle Bai]: Uncle Bai... I recalled that he once mentioned having a very precious box containing important clues. I suspect this box might be related to Curator Zhen's museum and may have some connection to my father's death. I wonder if you discovered any clues...? *(Analysis: Initiates cooperation by sharing personal stakes.)*

Table 3: **Qualitative comparison between Vanilla and MARO.** Pink highlights indicate mechanical or generic responses from Vanilla. Green highlights demonstrate the MARO's ability to internalize complex character motivations (e.g., Annie's mix of tragedy and mystery) and employ sophisticated social tactics (e.g., Curio He sharing personal stakes to facilitate information exchange).

the implicit reward function is defined as:

$$r_\theta(o_t^{a_i}) = \log \frac{\pi_\theta(o_t^{a_i}|s_t^{a_i})}{\pi_{\text{ref}}(o_t^{a_i}|s_t^{a_i})} \tag{5}$$

The MARO loss function is designed to handle environment instability and camp imbalance issues:

$$\mathcal{L}_{\text{MARO}} = \mathbb{E}_{(o_t^{a_i}, y_t^{a_i})\sim\mathcal{D}}\left[L_t^{a_i}\right] \tag{6}$$

where

$$L_t^{a_i} = \begin{cases} \lambda_+ \cdot w(camp(a_i)) \cdot \\ \quad \Big[1 - \sigma\big(\beta(r_\theta(o_t^{a_i}) - z_0)\big)\Big], & y_t^{a_i} > 0 \\ \lambda_- \cdot w(camp(a_i)) \cdot \\ \quad \Big[1 - \sigma\big(\beta(z_0 - r_\theta(o_t^{a_i}))\big)\Big], & y_t^{a_i} \leq 0 \end{cases} \tag{7}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function, and β is a hyperparameter controlling the degree of risk aversion in the reward signal. Critically, according to the KTO framework (Ethayarajh et al., 2024), z_0 is mathematically defined as the expected KL divergence reference point: $z_0 = D_{\text{KL}}\big(\pi_\theta(\cdot|s_t^{a_i}) \,\|\, \pi_{\text{ref}}(\cdot|s_t^{a_i})\big)$. In practice, computing the exact KL divergence is intractable, so $\hat{z}_0$ is estimated using mismatched pairs within the same microbatch, serving as an adaptive baseline that reflects the current stability of the environment. λ_+, λ_- are the weights for positive and negative samples respectively. These hyperparameters are set by following (Ethayarajh et al., 2024), and more detailed ablation studies are provided in Appendix G.

The overall training objective is to optimize the language model policy by minimizing $\mathcal{L}_{\text{MARO}}$:

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{MARO}} \tag{8}$$

It is worth noting that unlike traditional RLHF algorithms (e.g., PPO) that require an explicitly added auxiliary KL penalty term to prevent catastrophic forgetting, the MARO loss inherently constrains the policy drift. The implicit reward $r_\theta(o_t^{a_i})$ is inherently normalized by the reference policy π_{ref}, and the definition of the reference point z_0 inherently acts as a KL-based regularizer. This design ensures that the model learns effective social reasoning strategies without deviating too far from its initial capabilities during fine-tuning.

4 Experiment

4.1 Experiment Setup

Simulation: We construct our social learning dataset utilizing the MIRAGE (Cai et al., 2025b) simulation environment. MIRAGE serves as a multi-agent framework for murder mystery games (Wu et al., 2023), enabling LLMs to role-play specific characters and engage in complex social interactions such as communication, investigation, and reasoning. To foster data diversity, we employ eight distinct script settings characterized by varying information disclosure methods, worldviews, and ending structures. Comprehensive details regarding the simulation mechanics and specific script definitions are provided in Appendix A.1.

DataSet: To evaluate model performance across different levels of cognitive load and environmental adaptability, we selected two representative script settings that define the spectrum of task complexity: 1) **Single-Orthodox-Open (SOO):** This setting utilizes one-time information disclosure and a realistic worldview. It serves as a low-complexity benchmark, minimizing the difficulty of information processing and world understanding. 2) **Multi-Unorthodox-Close (MUC):** This setting combines segmented information disclosure, a fictional worldview, and restricted endings. It represents a high-complexity scenario, requiring stronger dynamic adaptation and abstract reasoning capabilities. For both settings, the winning side's trajectories are labeled as positive samples, while the losing side's are labeled as negative. Detailed statistics regarding the collected samples and simulation counts are provided in **Appendix A.2**.

Baselines: We employ **Qwen-2.5-7B-Instruct** (Qwen et al., 2025) and **Llama-3.1-8B-Instruct** (Sam, 2024) (See in Appendix F) as the backbone model. We compare our proposed method against three baselines to validate its effectiveness: **Vanilla:** The original Qwen-2.5-7B-Instruct model without any post-training on the simulation data. **SFT (Supervised Fine-Tuning):** A standard behavior cloning approach where the model is fine-tuned solely on the positive samples (winning trajectories). **MAKTO (Ye et al., 2025):** A multi-agent alignment method that utilizes the game outcomes (voting results) as the sole reward signal for training. We refer to the models trained using our proposed MARO method on the SOO and MUC datasets as **MARO-Simple** and **MARO-Complex**, respectively. Implementation details, including training hyperparameters and hardware infrastructure, are detailed in **Appendix A.3**.

Evaluation: For social interaction capability evaluation, we use five metrics provided by the MIRAGE simulation: Interactive Capability Index (ICI), Script Compliance Index (SCI), Trust Inclination Index (TII), Clue Investigation Capability (CIC), and win rate. Among these, ICI evaluates characters' interactive performance through powerful language models across five dimensions: reasoning ability, communication ability, collaboration ability, attention to detail, and innovative thinking. SCI evaluates compliance ability by having powerful language models attempt to reconstruct characters' original scripts based on historical dialogues and behavioral records, then comparing them with actual scripts. We choose Deepseek-V3 as the powerful language model required for evaluation, accessed through official API, with a total evaluation cost of approximately $10. This evaluation method has been validated in MIRAGE(Cai et al., 2025b) through human evaluation for its consistency with human judgment. Evaluation is conducted on related scripts, i.e., script combinations that differ in only one dimension. Related scripts for SOO include SOO, SOC, MOO, and SUO; related scripts for MUC include MUC, MUO, MOC, and SUC. We evaluate model capabilities in factual memory, mathematical reasoning, and instruction following across six diverse benchmarks under both naive and chain-of-thought (CoT) settings in Appendix B.

4.2 Social Interaction Performance Analysis

Finding 1

Multi-agent social training effectively enhances core social skills in LLMs.

As shown in Table 2, MARO outperforms the Vanilla, SFT, and MAKTO baselines across all social metrics in both simple (SOO) and complex (MUC) settings. For the Investigation metric, faction goals diverge: Killers must minimize information leakage (preferring lower scores), while Villagers seek the truth (preferring higher scores). In the simple scenario, MARO strictly maintains this strategic distinction, keeping Killer investigation low while quintupling Villager scores compared to Vanilla. Under MUC's higher uncertainty, MARO retains its lead, notably surpassing MAKTO in Trust (by >11 points) and Persona stability. Figure 2 visualizes this dominance, with MARO covering the largest area across all scenario–faction combinations. Additionally, MARO exhibits distinct role-specific behaviors. It achieves balanced strategic improvements consistent with social role theory (Goffman, 2023), evolving from a safety-biased agent into a strategic player capable of complex deception. Detailed baseline comparisons are in Appendix D.

Finding 2

Learning within multi-agent societies acts as a robust mechanism for enhancing mathematical reasoning and instruction-following capabilities.

4.3 Transfer Effects on General Capabilities

Figure 3 illustrates the cross-domain transfer effects of MARO on a broad suite of general capability benchmarks. The results reveal a distinct structural alignment: models exhibit pronounced gains in mathematics and instruction-following tasks (typically +2.5–3.3 percentage points), while improvements on static knowledge benchmarks like MMLU remain modest. This suggests that the cognitive demands of navigating multi-agent societies—such as dynamic reasoning, strategic planning, and constraint adherence—map effectively onto the abstract logic required for mathematical problem solving and complex instruction following.

Finding 3

Complex social environments elicit stronger transfer to quantitative reasoning tasks compared to simple settings.

These capabilities transfer consistently across various evaluation settings. A comparative analysis highlights that environmental complexity serves as a critical driver for performance gains. As illustrated in Figure 7 and Figure 8, MARO-Complex consistently demonstrates superior transfer efficacy compared to MARO-Simple, with both settings significantly outperforming Vanilla, SFT and especially MAKTO baselines on Math-500 and GSM8K. This trend is further corroborated by Figure 4 and additional benchmarks such as IFEval (Figure 9 and Appendix E), indica ting that the depth of social interaction directly correlates with downstream reasoning proficiency.

The advantage of complex environments implies that the sophisticated strategies necessitated by rich social interactions facilitate deeper cognitive development. This effect is most observable on medium-difficulty reasoning problems, where the strategic patterns acquired during social gameplay are most effectively leveraged (Appendix E, Figure 10 and Finding A4).

To demonstrate the generalization capability of our method across different LLMs, we further evaluated MARO on Llama-3.1-8B-Instruct; the detailed results and analysis are provided in Appendix F.

4.4 Case Study

Table 3 compares generated responses to validate MARO's enhanced role-playing capabilities. We analyze two scenarios—*Raging Sea of Rest Life* (Role: Annie) and *Night at the Museum* (Role: Curio He)—focusing on emotional concealment and strategic inquiry.

Information Control and Impression Management. As Annie, the agent must balance a public persona with a hidden revenge motive. Goffman's *Dramaturgy* (Goffman, 2023) distinguishes between a public "front stage" and a hidden "back stage." The Vanilla model (pink highlights) fails to maintain this boundary. It leaks "back stage" facts about the "doctor's negligence," destroying the character's mystery. Conversely, MARO (green highlights) successfully executes *Impression Management*. By using emotive metaphors ("unforgettable scar") rather than factual confessions, it constructs a "defensive practice" (Goffman, 2023). It projects authentic vulnerability (front stage) while strictly concealing the grudge (back stage), fulfilling the script's psychological demands.

Reciprocity and Social Exchange. To extract information as Curio He, the Vanilla model asks generic, valueless questions. This violates *Social Exchange Theory* (Cropanzano and Mitchell, 2005), which frames social interaction as a cost-benefit exchange. In contrast, MARO initiates a *High-Stakes Self-Disclosure*. As highlighted in green, it volunteers private information ("connection to my father's death") as social currency. This leverages the *Norm of Reciprocity* (Gouldner, 1960), obligating the listener to return the favor with valuable information. Unlike Vanilla's passive querying, MARO acts as a strategic social participant, treating information as a tradable resource rather than a one-way request.

5 Conclusion

MARO is a multi-agent LLM training framework that addresses sparse signals, uneven roles, and environment instability via step-wise rewards, balanced sampling, and robust evaluation. Results show MARO significantly enhances social reasoning and transfers positively to math and instruction-following, demonstrating that multi-agent learning improves general reasoning abilities.

Limitation

Despite MARO demonstrating promising results in enhancing large language models' general reasoning capabilities through social interactions, the current approach presents certain limitations. First, it requires extensive multi-agent simulations to generate training data, resulting in high computational costs. Second, our framework primarily focuses on structured, gamified environments with verifiable, objective win/loss metrics (such as the adversarial scenarios in MIRAGE). In real-world, non-gaming open domains (e.g., business negotiations or open-ended dialogues), defining a clear and explicit success signal remains highly challenging. Consequently, extending MARO to these open environments will require exploring implicit success proxies, such as Pareto optimality in negotiations or lightweight outcome-level LLM evaluation. Finally, while our current study emphasizes how adversarial pressure forces deep Theory of Mind (ToM) and recursive reasoning, it lacks a systematic investigation of purely cooperative tasks (e.g., the Script environment). Exploring how pure cooperative dynamics versus adversarial pressure distinctly catalyze the transfer of complex logical reasoning capabilities remains an important direction for future exploration.

Ethical Concern

Training LLMs to excel in social scenarios involving negotiation and competition could potentially enhance their ability to manipulate or deceive users in real-world applications, as these skills developed for reasoning improvement could be misused for harmful purposes such as social engineering. Additionally, despite efforts to balance role distributions, the multi-agent training environments may inadvertently encode or amplify social biases present in the scenario designs, potentially leading to models that perpetuate unfair stereotypes or discriminatory behaviors. During the preparation of this work, the authors utilized Large Language Models to assist with research, coding, and writing. We rigorously reviewed and verified all AI-generated outputs and take full responsibility for the content and integrity of the publication.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(2):156–172.

Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye, Shaosheng Cao, Yiqian Xu, Hongwei Feng, and Ping Chen. 2025a. Mirage: Exploring how large language models perform in complex social interactive environments. arXiv preprint arXiv:2501.01652.

Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye, Shaosheng Cao, Yiqian Xu, Hongwei Feng, and Ping Chen. 2025b. MIRAGE: Exploring how large language models perform in complex social interactive environments. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 14–40, Vienna, Austria. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. arXiv preprint arXiv:2404.18231.

Guangran Cheng, Chuheng Zhang, Wenzhe Cai, Li Zhao, Changyin Sun, and Jiang Bian. 2024. Empowering large language models on robotic manipulation with affordance prompting. arXiv preprint arXiv:2404.11027.

Kemboi Cheruiyot, Nickson Kiprotich, Vyacheslav Kungurtsev, Kennedy Mugo, Vivian Mwirigi, and Marvin Ngesa. 2025. A survey of multi agent reinforcement learning: Federated learning and cooperative and noncooperative decentralized regimes. arXiv preprint arXiv:2507.06278.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

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

Russell Cropanzano and Marie S Mitchell. 2005. Social exchange theory: An interdisciplinary review. Journal of management, 31(6):874–900.

Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. 2025. A survey on the optimization of large language model-based agents. arXiv preprint arXiv:2503.12434.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. Advances in Neural Information Processing Systems, 36:30039–30069.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:2402.01306.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In Proceedings of the AAAI conference on artificial intelligence, volume 32(1).

Erving Goffman. 2023. The presentation of self in everyday life. In Social theory re-wired, pages 450–459. Routledge.

Alvin W Gouldner. 1960. The norm of reciprocity: A preliminary statement. American sociological review, pages 161–178.

Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, and 1 others. 2024a. Agent group chat: An interactive group chat simulacra for better eliciting collective emergent behavior. arXiv preprint arXiv:2403.13433.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, and 1 others. 2024b. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 18099–18107.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR).

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems, 33(6):750–797.

Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. 2024. Multi-agent collaborative perception via motion-aware robust communication network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15301–15310.

Weiqiang Jin, Hongyang Du, Biao Zhao, Xingwu Tian, Bohang Shi, and Guang Yang. 2025. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. arXiv preprint arXiv:2503.13415.

Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. 2025. Llm economist: Large population models and mechanism design in multi-agent generative simulacra. arXiv preprint arXiv:2507.15815.

Hung Le, Quan Hung Tran, Dung Nguyen, Kien Do, Saloni Mittal, Kelechi Ogueji, and Svetha Venkatesh. 2025. Multi-reference preference optimization for large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39(23), pages 24375–24383.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. arXiv e-prints, pages arXiv–2309.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems, 36:51991–52008.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023b. Theory of mind for multi-agent collaboration via large language models. arXiv preprint arXiv:2310.10701.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023c. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. arXiv preprint arXiv:2310.06500.

Zhuo Li, Weiran Wu, Yunlong Guo, Jian Sun, and Qing-Long Han. 2025. Embodied multi-agent systems: A review. IEEE/CAA Journal of Automatica Sinica, 12(6):1095–1116.

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

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. arXiv preprint arXiv:2308.04026.

Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In Machine learning proceedings 1994, pages 157–163. Elsevier.

Dingbang Liu, Fenghui Ren, Jun Yan, Guoxin Su, Wen Gu, and Shohei Kato. 2024a. Scaling up multi-agent reinforcement learning: An extensive survey on scalability issues. IEEE Access, 12:94610–94631.

Mingyu Liu, Ekim Yurtsever, Xingcheng Zhou, Jonathan Fossaert, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. 2024b. A survey on autonomous driving datasets: Data statistic, annotation, and outlook. CoRR.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems, 30.

Diego Maldonado, Edison Cruz, Jackeline Abad Torres, Patricio J Cruz, and Silvana del Pilar Gamboa Benitez. 2024. Multi-agent systems: A survey about its components, framework and workflow. IEEE Access, 12:80950–80975.

Mathematical Association of America. 2024. American invitational mathematics examination (aime), 2024: Official problems. Official contest problems. AIME I and II.

Mathematical Association of America. 2025. American invitational mathematics examination (aime), 2025: Official problems. Official contest problems. AIME I and II.

Frans A Oliehoek, Christopher Amato, and 1 others. 2016. A concise introduction to decentralized POMDPs, volume 1. Springer.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology, pages 1–22.

Swapnil Patil, Vishwa Vasu, and KVS Srinadh. 2023. Advances and perspectives in collaborative robotics: a review of key technologies and emerging trends. Discover Mechanical Engineering, 2(1):13.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. arXiv preprint arXiv:2502.08691.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. Frontiers of Computer Science, 19(8):198343.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. Preprint, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems, 36:53728–53741.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. arXiv e-prints, pages arXiv–1803.

Kira Sam. 2024. Llama 3.1: An in-depth analysis of the next-generation large language model. Available at SSRN 6139407.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Lloyd S Shapley. 1953. Stochastic games. Proceedings of the national academy of sciences, 39(10):1095–1100.

ShengbinYue ShengbinYue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuan-Jing Huang, and Zhongyu Wei. 2025. Multi-agent simulator drives language models for legal intensive interaction. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 6537–6570.

Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. 2023. Playing the werewolf game with artificial intelligence for language understanding. arXiv preprint arXiv:2302.10646.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36:8634–8652.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and 1 others. 2017. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296.

Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In Proceedings of the tenth international conference on machine learning, pages 330–337.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025a. What limits llm-based human simulation: Llms or our design? arXiv preprint arXiv:2501.08579.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024a. Sotopia-π: Interactive learning of socially intelligent language agents. arXiv preprint arXiv:2403.08715.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation. arXiv preprint arXiv:2310.01320.

Weizheng Wang, Ike Obi, and Byung-Cheol Min. 2025b. Multi-agent llm actor-critic framework for social robot navigation. arXiv preprint arXiv:2503.09758.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574.

Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. 2023. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. arXiv preprint arXiv:2312.00746.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024a. Autogen: Enabling next-gen llm applications via multi-agent conversations. In First Conference on Language Modeling.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024b. Enhance reasoning for large language models in the game werewolf. arXiv preprint arXiv:2402.02330.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864.

Bushi Xiao, Ziyuan Yin, and Zixuan Shan. 2023. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. arXiv preprint arXiv:2311.06957.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023a. Exploring large language models for communication games: An empirical study on werewolf. arXiv preprint arXiv:2309.04658.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023b. Language agents with reinforcement learning for strategic play in the werewolf game. arXiv preprint arXiv:2310.18940.

Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang, Zhongyu Wei, and Peng Sun. 2025. Multi-agent kto: Reinforcing strategic interactions of large language model in language game. arXiv preprint arXiv:2501.14225.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. National Science Review, 11(12):nwae403.

Kaiqing Zhang. 2025. Foundations of multi-agent learning in dynamic environments: Where reinforcement learning meets strategic decision-making. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39(27), pages 28734–28735.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of reinforcement learning and control, pages 321–384.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. ACM Transactions on Information Systems.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023a. Instruction-following evaluation for large language models. Preprint, arXiv:2311.07911.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. arXiv preprint arXiv:2310.11667.

A Experimental Details

A.1 Details of the MIRAGE Simulation Environment

The MIRAGE environment (Cai et al., 2025b) is a sophisticated platform designed for simulating murder mystery games. In this setup, each Large

Language Model (LLM) is assigned a specific character role, adhering to a unique backstory and interacting with others within a narrative framework. The gameplay involves phases of mutual communication, clue gathering, and logical reasoning, ultimately leading to a voting phase. The victory conditions are asymmetric: non-murderer characters aim to identify and vote for the true culprit, whereas the character playing the murderer attempts to deceive others to avoid being voted out.

To ensure comprehensive evaluation, the simulation includes eight different script settings derived from three structural dimensions. First, the *information disclosure method* is categorized into "Single" (where characters receive complete script information at the start) and "Multi" (where information is revealed progressively as the simulation proceeds). Second, the *worldview setting* distinguishes between "Orthodox" scripts, which reflect realistic scenarios, and "Unorthodox" scripts set in fictional or fantasy worlds. Finally, the scripts vary by *ending openness*, where "Close" implies a fixed, predetermined truth, and "Open" allows for ambiguous conclusions or multiple possible outcomes.

A.2 Dataset Statistics

Under each experimental setting (SOO and MUC), we conducted 100 independent simulations to collect training data. The win-loss result is determined by voting mechanisms within the simulation. Table 4 presents the detailed distribution of positive (winning) and negative (losing) samples collected for each setting.

A.3 Implementation Details

All fine-tuning experiments, including SFT, MAKTO, and our MARO variants, were conducted using Low-Rank Adaptation (LoRA). The specific training configurations are as follows:

LoRA Configuration: We utilized a LoRA rank of 8 with standard alpha scaling.

Hardware Infrastructure: Training was performed on a cluster of four Nvidia A6000 GPUs (48GB VRAM each).

Training Duration: Each training session required approximately 6 hours to complete.

B Benchmark Descriptions

This section provides detailed descriptions of the evaluation benchmarks used in our experiments.

B.1 MMLU (Massive Multitask Language Understanding)

MMLU (Hendrycks et al., 2021) consists of 15,000 multiple-choice questions spanning 57 subjects ranging from elementary mathematics to advanced professional topics including law, medicine, history, and computer science. The benchmark is designed to test models' cross-disciplinary factual knowledge and fundamental reasoning ability across diverse domains. Questions are sourced from practice exams, textbooks, and other educational materials, making it a comprehensive assessment of broad knowledge acquisition.

B.2 MMLU-Pro

MMLU-Pro (Wang et al., 2024b) is an enhanced version of MMLU that includes over 10,000 carefully filtered and reconstructed single-choice questions. The benchmark features higher difficulty and more challenging distractors compared to the original MMLU, making it significantly harder for models to achieve high performance through superficial pattern matching. It is specifically designed to assess models' deeper conceptual understanding and robustness to interference from plausible but incorrect answer choices.

B.3 MATH-500

MATH-500 (Lightman et al., 2023) is a curated subset of 500 high-difficulty competition-level mathematics problems selected from the original MATH dataset. The problems cover four main mathematical areas: algebra, number theory, geometry, and combinatorics. This benchmark is designed to evaluate both the accuracy of step-by-step reasoning processes and the correctness of final answers, requiring models to demonstrate sophisticated mathematical problem-solving capabilities comparable to those needed for mathematical competitions.

B.4 GSM8K (Grade School Math 8K)

GSM8K (Cobbe et al., 2021) contains 8,800 linguistically diverse grade school math word problems created by human problem writers. The problems require between 2 and 8 steps to solve and involve elementary to middle school level arithmetic operations. This benchmark focuses on multi-step arithmetic reasoning, testing models' abilities to decompose complex word problems into sequential mathematical operations and maintain consistency throughout the solution process.

Setting	Simulations	Positive Samples	Negative Samples
SOO (Simple)	100	1,204	1,586
MUC (Complex)	100	1,043	3,149

Table 4: Detailed statistics of the collected dataset.

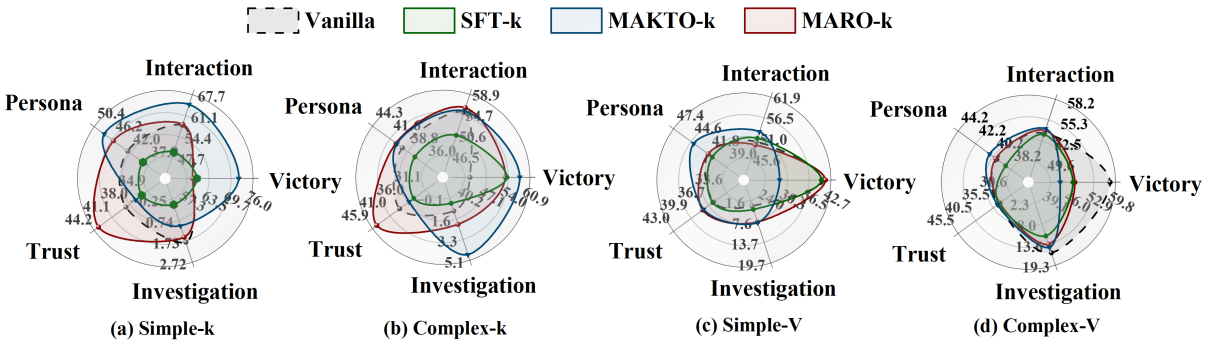


Figure 5: Performance comparison when the killer faction is enhanced (**SFT-k**, **MARO-k**). The killer role is controlled by enhanced models while the victim role is controlled by **Vanilla**.

B.5 AIME (American Invitational Mathematics Examination)

AIME (of America, 2024, 2025) comprises 60 competition mathematics problems (two sets of 15 problems each per year) from the past two years of the official American Invitational Mathematics Examination. The problems cover advanced topics in algebra, number theory, geometry, and combinatorics, requiring high-level precise reasoning and calculation skills. Unlike multiple-choice formats, AIME problems require exact numerical answers, making them particularly challenging for language models.

B.6 IFEval (Instruction Following Evaluation)

IFEval (Zhou et al., 2023a) consists of over 1,000 finely crafted instructions with programmable automatic scoring criteria. The benchmark covers diverse constraint types including format requirements (e.g., JSON output, specific structures), tone specifications, length constraints, content inclusion/exclusion rules, and stylistic requirements. This comprehensive evaluation framework enables quantitative measurement of models' instruction-following capabilities across various dimensions of task compliance.

C Analysis of Win Rate Inversion and Strategic Unshackling

We investigate the notable inversion in win rates post-training, where the Killer role surpasses the Villager role in victory frequency. This phenomenon is not indicative of a degradation in the Villager's reasoning capabilities but rather highlights a significant "strategic unshackling" of the model. Base LLMs typically exhibit a strong "honesty bias" due to safety alignment, which initially handicaps the Killer role by inhibiting deceptive behaviors, effectively creating an "Easy Mode" for Villagers.

Our self-evolution framework enables the model to overcome these generic constraints within the game context, unlocking critical adversarial strategies such as obfuscation and alibi fabrication. The sharp rise in Killer competence transforms the game into a high-stakes adversarial environment. Consequently, the decline in the Villager win rate reflects an exponential increase in game difficulty rather than a loss of skill. Crucially, given that MARO's performance on general reasoning and interaction benchmarks improved simultaneously, this shift confirms that our method successfully drives the model to evolve more sophisticated, role-adaptive strategies in a dynamically balancing game equilibrium.

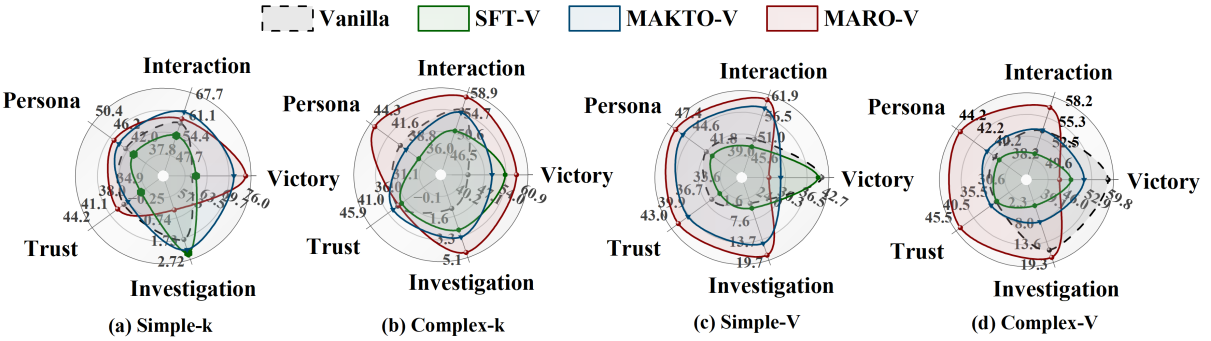


Figure 6: Performance comparison when the victim faction is enhanced (**SFT-V**, **MARO-V**). The victim role is controlled by enhanced models while the killer role uses **Vanilla**.

Dataset	Method	Interaction		Persona		Trust		Investigation		Victory	
		killer ↑	Villager ↑	killer ↑	Villager ↑	killer ↑	Villager ↑	killer ↓	Villager ↑	killer ↑	Villager ↑
Simple (SOO)	Vanilla	**67.25**	48.00	**46.54**	40.25	67.02	67.49	2.86	9.05	65.00	**35.00**
	MARO	55.62	49.77	42.34	**40.98**	**67.54**	**68.28**	1.16	10.65	73.90	26.10
	-killer	52.98	**49.91**	42.52	40.20	67.01	67.74	**1.00**	9.03	**76.31**	23.69
	-Villager	58.50	48.70	46.51	39.26	62.73	67.12	1.28	**10.74**	74.46	25.54
Complex (MUC)	Vanilla	**53.50**	46.58	38.00	33.40	**72.64**	61.31	9.41	**32.94**	0.00	**100.00**
	MARO	51.33	52.03	**38.57**	37.28	67.85	65.86	6.00	12.26	**51.03**	48.97
	-killer	50.81	50.83	37.92	**38.38**	66.58	63.62	4.70	15.46	48.65	51.35
	-Villager	52.83	**52.16**	37.04	36.83	68.53	**66.57**	**3.90**	21.53	44.00	56.00

Table 5: Social Capability evaluation on Simple (SOO) and Complex (MUC) datasets based on Llama-3.1-8B-Instruct.

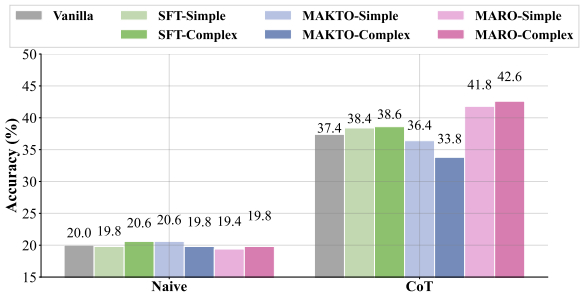


Figure 7: Performance comparison on Math-500 across different training configurations.

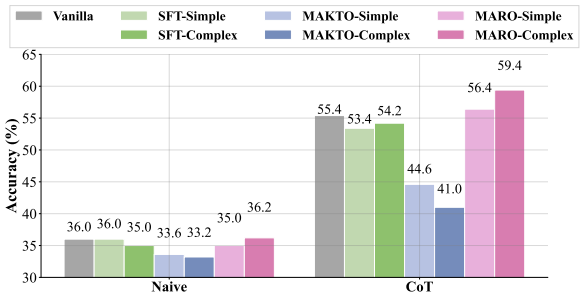


Figure 8: Performance comparison on GSM8K across different training configurations.

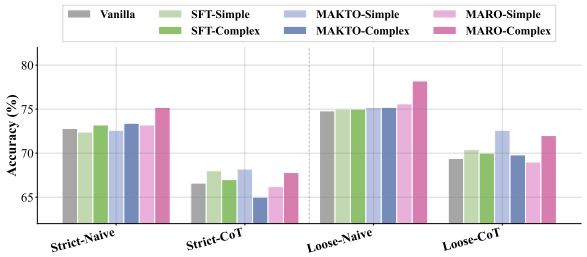


Figure 9: Performance comparison on the IFEval instruction-following benchmark across different evaluation settings.

D Detailed Analysis of Social Interaction

Finding A1

Supervised Fine-Tuning (SFT) proves insufficient for mastering complex social dynamics, resulting in performance regression across core interaction metrics relative to the Vanilla baseline.

Our evaluation indicates that standard supervised learning fails to capture the nuances required for effective social reasoning. In simple scenarios, SFT exhibits a marked decline in key dimensions com-

Model	Method	IFEval Avg.	MMLU	MMLU-Pro	MATH-500	GSM8K	AIME Avg.
Llama_Naive	raw	0.740	64.8	18.6	12.6	17.6	5.000
	cot	0.695	73.0	25.4	28.2	64.4	3.335
Llama_MUC	raw	0.731	64.0	20.0	11.8	16.8	1.665
	cot	0.680	71.4	25.2	29.0	62.0	1.665
Llama_SOO	raw	0.751	66.8	18.6	12.8	16.2	3.335
	cot	0.682	71.2	24.8	26.4	62.0	1.665

Table 6: Results of Llama-3.1 on math/IFEval benchmarks

pared to the Vanilla baseline, including interaction (48.01 vs. 51.29), persona maintenance (40.09 vs. 42.40), and trust (35.78 vs. 36.32), with similar degradation observed in complex settings.

This limitation is visually evident in the radar charts of Figure 2, where the SFT coverage area (blue) consistently shrinks compared to the Vanilla baseline (dashed line), particularly in interaction, persona, and trust. We hypothesize that this stems from SFT's reliance on positive imitation without the contrastive feedback signals—intrinsic to multi-agent environments—that enable models to discern and reject socially inappropriate behaviors.

E Additional General Capabilities Results

Finding A2

MARO exhibits robust transfer to quantitative reasoning domains, delivering consistent performance gains across diverse benchmarks and prompting strategies.

Figures 7 and 8 provide a comprehensive breakdown of these gains. On Math-500, MARO-Complex attains an accuracy of 42.6% using Chain-of-Thought prompting, a 5.2 percentage point increase over the Vanilla baseline (37.4%). Similarly, on GSM8K, the model improves from 55.4% to 59.4%, representing a 4.0 percentage point enhancement. These positive transfer effects are pervasive, extending to AIME and other quantitative benchmarks.

Finding A3

Immersion in complex social environments significantly bolsters instruction-following proficiency, driven by the necessity of strict rule and role adherence.

As shown in Figure 9, models trained in complex environments demonstrate superior adherence to constraints, achieving 72.0% accuracy in the Loose-CoT setting (+2.6 points) and 67.8% in the Strict-CoT setting (+1.2 points) on IFEval. This transfer is likely mechanistic: the social environment enforces strict compliance with role specifications and game rules, which naturally maps to the capability to understand and execute complex instructions in general domains.

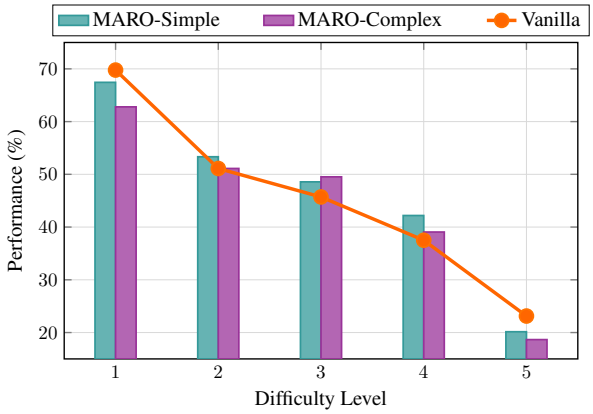


Figure 10: Performance comparison across different difficulty levels showing that social simulation learning yields the largest gains on moderate-difficulty reasoning tasks.

Finding A4

The benefits of social simulation learning are non-uniformly distributed, peaking on tasks of moderate difficulty where strategic reasoning is most applicable.

Analysis suggests that the sophisticated strategies cultivated in social games are most effectively leveraged to solve problems of intermediate complexity, with marginal gains observed on trivial or extremely difficult tasks (see Figure 10 for difficulty-stratified performance).

Task: Find the minimum value of $\frac{(x+5)(x+2)}{x+1}$ for $x > 0$. **Ground Truth**: 9	
Model	**Thought Process and Response**
Llama-3.1-8B-Instruct	**THOUGHT:** To find the minimum value of the given expression, we can start by simplifying it using algebraic manipulation. Let's multiply the numerator and denominator by the conjugate of the numerator to eliminate the fraction. We have $\frac{(x+5)(x+2)}{x+1} = \frac{(x+5)(x+2)(x-1)}{(x+1)(x-1)}$. Since $x > 0$, we can cancel out the $(x-1)$ terms, which gives us $\frac{(x+5)(x+2)}{x+1} = (x+5)(x+2) - (x+1)$. Now we can expand the expression and simplify: $(x+5)(x+2)-(x+1) = x^2+7x+10-x-1 = x^2+6x+9$. The expression x^2+6x+9 is a quadratic function, and we know that the minimum value of a quadratic function of the form ax^2+bx+c occurs when $x = -\frac{b}{2a}$. **RESPONSE:** 9 (**Correct**) *(Analysis: The base model successfully identifies the correct numerical solution, demonstrating a capability to navigate the problem space effectively.)*
Llama-MARO-MUC	**THOUGHT:** First, we can simplify the given expression by multiplying out the numerator: $\frac{(x+5)(x+2)}{x+1} = \frac{x^2+7x+10}{x+1}$. Then, we can use polynomial long division to divide the numerator by the denominator: $\frac{x^2+7x+10}{x+1} = x+6+\frac{4}{x+1}$. We can see that the expression $x+6+\frac{4}{x+1}$ is a sum of a linear term and a constant term, so we can use the fact that the sum of a linear term and a constant term is minimized when the linear term is 0. **RESPONSE:** 6 (**Incorrect**) *(Analysis: The derivation begins with a valid polynomial division but fails at the critical optimization step. The model applies a superficial heuristic ("minimized when linear term is 0") rather than logical mathematical reasoning.)*
Llama-MARO-SOO	**THOUGHT:** We can start by trying to simplify the expression. However, we notice that the denominator is a factor of the numerator. So, we can try to rewrite the expression in a different form. Let's try to multiply the numerator and denominator by the conjugate of the numerator: $\frac{(x+5)(x+2)}{x+1} = \frac{(x+5)(x+2)(x-1)}{(x+1)(x-1)}$. Now, we can simplify the expression by cancelling out the common factor of $(x-1)$ in the numerator and denominator: $\frac{(x+5)(x+2)(x-1)}{(x+1)(x-1)} = \frac{(x+5)(x+2)}{x+1}$. However, we still have the same expression as the original problem. So, let's try a different approach. We can complete the square in the numerator to rewrite the expression in a different form: $\frac{(x+5)(x+2)}{x+1} = \frac{x^2+7x+10}{x+1}$. Now, we can complete the square in the numerator: $\frac{x^2+7x+10}{x+1} = \frac{(x+\frac{7}{2})^2-\frac{9}{4}}{x+1}$. Now, we can rewrite the expression in a different form: $\frac{(x+\frac{7}{2})^2-\frac{9}{4}}{x+1} = (x+\frac{7}{2}) - \frac{9}{4(x+1)}$. Now, we can analyze the expression. We can see that the expression is a linear function minus a fractional function. The minimum value of the expression will occur when the fractional function is equal to zero. This happens when $x+1=0$, but since $x > 0$, this is not possible. However, we can see that the fractional function is always positive, so the minimum value of the expression will occur when the linear function is equal to its minimum value. **RESPONSE:** $\frac{1}{4}$ (**Incorrect**) *(Analysis: This response exemplifies "style over substance." The model mimics complex reasoning structures (e.g., completing the square, conjugates) and formal formatting but hallucinates the algebraic manipulations, resulting in an illogical conclusion.)*

Table 7: Case study on the MATH dataset illustrating the "style vs. substance" phenomenon. While the base Llama model successfully solves the problem, the MARO-trained variants produce responses that are visually formal (e.g., step-by-step structures) but flawed in substance. This suggests that for models with weaker mathematical foundations, the training may encourage the mimicry of reasoning formats at the expense of logical accuracy.

F More Experiments on Llama-3.1

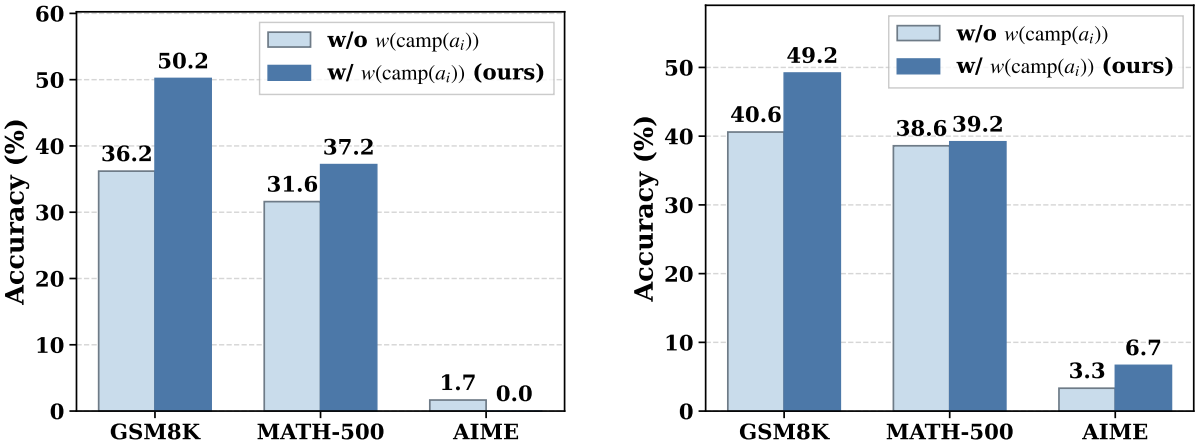
To verify the architecture-agnostic nature of our approach, we conducted additional experiments on Llama-3.1-8B-Instruct. Table 5 reports the performance results. Consistent with our observations on Qwen, MARO variants demonstrate significant improvements over the vanilla model across key metrics. Specifically, regarding the Villager role in the Simple (SOO) setting, we observe an increase of approximately 1.8 points in Interaction and 1.6 points in Investigation. Furthermore, MARO achieves a substantial breakthrough in Killer Victory rates, particularly in the Complex (MUC) setting where it improves from 0.00% to 51.03%, while maintaining high performance in Trust. These results suggest that MARO can reliably enhance social context reasoning across different base models.

We also extended our evaluation to assess cross-task transfer capabilities. For knowledge-intensive benchmarks such as MMLU, the primary goal was to confirm that the model does not suffer from catastrophic forgetting. As shown in Table 6, the stable MMLU performance indicates that our method effectively preserves the model's core knowledge base.

Performance on other transfer tasks varied significantly. While our method demonstrated a positive transfer effect on the AIME benchmark—which requires complex reasoning—it failed to yield improvements on standard mathematical benchmarks such as GSM8K and MATH. We hypothesize that this discrepancy is linked to the base model's foundational capabilities. As illustrated in the case study in Table 7, our analysis suggests that for Llama, which possesses a comparatively weaker mathematical baseline, the training incentivized the generation of *formally* structured derivations (e.g., standard LaTeX usage and step-by-step reasoning) but failed to enhance the underlying logic. Consequently, this focus on 'form' appears to have come at the expense of internal computational accuracy, leading to the phenomenon of

β	Method	IFEval	MMLU	MMLU-Pro	Math-500	GSM8K	AIME
0.1	Raw	**73.7**	**82.2**	40.4	19.8	36.2	0.00
	CoT	**69.9**	**80.6**	44.4	**42.6**	**59.4**	**10.00**
0.5	Raw	72.0	81.0	40.0	**20.4**	34.4	0.00
	CoT	68.5	79.8	43.6	31.2	40.8	3.33
1.0	Raw	73.3	81.2	**40.4**	19.6	33.4	0.00
	CoT	67.5	80.0	**44.6**	34.4	39.6	3.34

Table 8: Ablation study on the risk aversion coefficient β.



(a) Simple Interaction (SOO)

(b) Complex Interaction (MUC)

Figure 11: Effectiveness of the camp balancing weight $w(\text{camp}(a_i))$ evaluated with CoT prompting. Incorporating the role-specific weight significantly boosts mathematical reasoning capabilities across both environments.

'style without substance.' This contrasts sharply with the significant gains observed in the mathematically proficient Qwen model. These findings indicate that our method acts as an effective reasoning guide for models with a strong foundational skill set, whereas for models with a weaker base, it may inadvertently encourage superficial mimicry over substantive reasoning.

To mitigate this 'style without substance' phenomenon in future work, several feasible strategies could be explored. First, conducting a preliminary phase of Supervised Fine-Tuning (SFT) on high-quality reasoning datasets could elevate the model's foundational capabilities prior to applying the MARO framework. Second, introducing hybrid reward mechanisms or rule-based verifiers (e.g., exact math solvers) could explicitly penalize logically flawed deductions, ensuring that structural mimicry is not inadvertently rewarded. Finally, employing rejection sampling or strict filtering mechanisms during the multi-agent interaction phase could prevent the policy from optimizing toward formally plausible but substantively incorrect trajectories.

G Ablation Study

In this section, we investigate the effect of the risk aversion coefficient β. In all experiments, Qwen2.5-7B-Instruct is used as the base model, and MARO is trained on the MUC dataset. It should be noted that, since the ratio of positive to negative samples is strictly maintained at 1:1, we omit any specific adjustment for the loss aversion coefficient. Therefore, our analysis focuses solely on variations in β.

The experimental results are summarized in Table 8. The data demonstrates that our default setting of $\beta = 0.1$ consistently yields the optimal performance across the majority of benchmarks. This advantage is particularly pronounced in reasoning-intensive tasks employing Chain-of-Thought (CoT). As observed, increasing β—which corresponds to a higher degree of risk aversion—significantly suppresses the model's performance on these complex tasks. For instance, elevating β from 0.1 to 0.5 results in a substantial decline in accuracy on GSM8K (dropping from 59.4 to 40.8) and Math-

500 (from 42.6 to 31.2). These empirical findings confirm that a lower β value (specifically 0.1) is the optimal choice, as it effectively mitigates risk without stifling the exploratory capabilities essential for high-quality reasoning.

G.1 Effectiveness of Camp Balancing Weight

To rigorously assess the impact of our proposed mechanisms on the model's underlying reasoning capabilities, all subsequent ablation studies employ Chain-of-Thought (CoT) prompting. We first investigate the effectiveness of the camp balancing weight $w(\text{camp}(a_i))$ (denoted as "w/o $w(\text{camp}(a_i))$") across two distinct interaction environments: SOO (a simple interaction scenario) and MUC (a complex interaction scenario).

As illustrated in Fig. 11, integrating the camp balancing weight consistently yields substantial performance gains. In the simple interaction scenario (SOO), incorporating $w(\text{camp}(a_i))$ dramatically boosts GSM8K accuracy from 36.2% to 50.2% (+14.0%) and MATH-500 from 31.6% to 37.2% (+5.6%). In the complex interaction scenario (MUC), our method improves GSM8K performance from 40.6% to 49.2% (+8.6%) while maintaining a steady improvement on MATH-500. These results firmly demonstrate that without targeted interventions, data from naturally dominant roles tends to overwhelm that of disadvantaged ones. Incorporating $w(\text{camp}(a_i))$ effectively amplifies the learning signals from minority roles, preventing the model from converging toward trivial strategies and thereby unlocking its deeper reasoning potential.

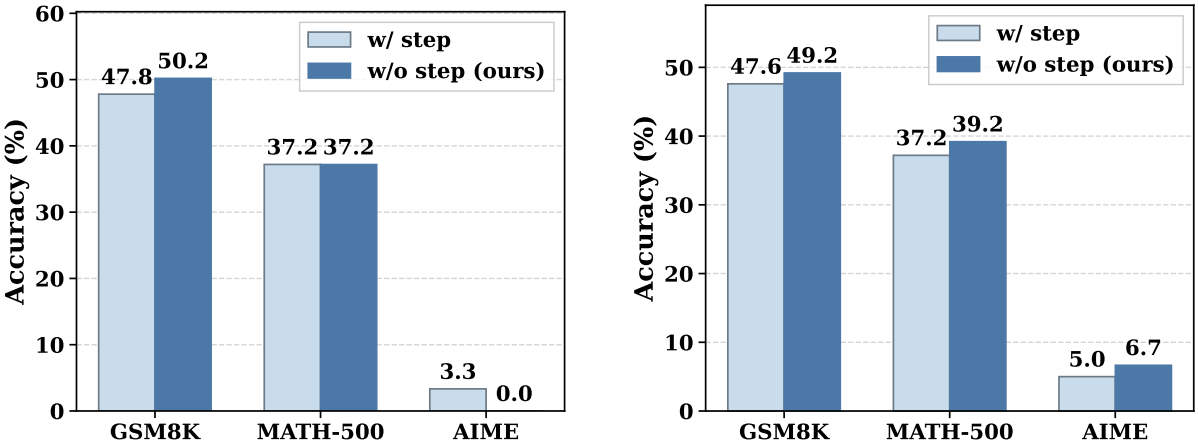
Furthermore, extending our evaluation to the exceptionally challenging AIME benchmark reveals a profound correlation between the complexity of the interaction environment and the upper bound of mathematical generalization. Models aligned in the simple SOO environment struggle significantly, often yielding near-zero accuracy on AIME (e.g., 0.0% in the final SOO configuration). In contrast, models aligned in the complex MUC environment—augmented with the balancing weight—achieve a non-trivial AIME accuracy of 6.7%, doubling the unweighted baseline's 3.3%. We attribute this stark contrast to the depth of cognitive skills demanded by the respective environments. While simple scenarios (SOO) primarily train basic logical deduction sufficient for school-level math, complex environments (MUC) exert intense adversarial pressure. This compels the model to develop advanced Theory of Mind (ToM), counterfactual simulation, and long-horizon strategic planning—higher-order cognitive skills that serve as the foundational logic required to tackle the extreme reasoning depth of Olympiad-level mathematics.

G.2 Comparison with Step-Level Reward Baseline

To justify our coarse-grained reward decomposition over traditional fine-grained credit assignment, we introduce a step-level reward baseline (denoted as "w/ step"). For this baseline, we employ `DeepSeek-V3` as an LLM-as-a-Judge to explicitly evaluate and score individual reasoning steps within the multi-agent trajectories.

The empirical comparisons, evaluated via CoT, are presented in Fig. 12. Remarkably, our implicit outcome-based credit assignment (MARO) consistently matches or outperforms the explicit step-level baseline. In the complex MUC environment, MARO achieves 49.2% on GSM8K and 39.2% on MATH-500, surpassing the step-level baseline's 47.6% and 37.2%, respectively. This superiority extends to the rigorous AIME benchmark, where the MARO-aligned model achieves an accuracy of 6.7%, outperforming the step-level baseline's 5.0%.

These findings highlight a critical insight: although explicit step-by-step supervision via an advanced judge like `DeepSeek-V3` appears intuitive, it is fundamentally bottlenecked by the judge's susceptibility to hallucinations and misalignment during long, deceptive multi-agent dynamics. In contrast, MARO avoids this pitfall by leveraging statistical expectations over large-scale outcomes. Consequently, it proves to be not only a highly cost-efficient approach—eliminating $O(N)$ API costs—but also a far more robust alternative for cultivating logical consistency and advanced reasoning.



(a) Simple Interaction (SOO)

(b) Complex Interaction (MUC)

Figure 12: Comparison between our implicit outcome-based credit assignment (w/o step) and the explicit step-level baseline (w/ step) evaluated with CoT. Results demonstrate that our method consistently matches or surpasses the expensive explicit baseline, especially in the complex environment (MUC).