

# 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

084	only to understand language but also to learn how	ducted extensive experiments in simulated social	136
085	to make wise decisions in dynamically changing	environments. The results demonstrate that MARO	137
086	environments, handle conflicts of interest, and even	significantly enhances large language models' rea-	138
087	change strategies when necessary to adapt to new	soning abilities across multiple dimensions, with	139
088	situations (Jin et al., 2025; Maldonado et al., 2024).	improvements in social reasoning, transferable	140
089	However, training LLMs in multi-agent environ-	gains to mathematical reasoning and instruction-	141
090	ments faces three key challenges (Liu et al., 2024a).	following tasks, and superior performance com-	142
091	First is the sparse learning signal problem: for in-	pared to traditional supervised fine-tuning methods.	143
092	dividual LLMs, during long interaction processes,	Importantly, complex social environments proved	144
093	only the final success or failure results are clear, but	more effective than simple ones in promoting gen-	145
094	LLMs need to know which specific decisions dur-	eral reasoning capability enhancement.	146
095	ing the lengthy interaction process were effective		
096	and which were harmful (Cai et al., 2025a). Second	<b>2 Related Work</b>	147
097	is the uneven role distribution problem: in many	<b>2.1 Traditional Multi-Agent Systems and</b>	148
098	social scenarios, the difficulty of success varies	<b>Reinforcement Learning</b>	149
099	greatly among different roles. For example, in busi-	Traditional Multi-Agent Systems originate from	150
100	ness negotiations, sellers often find it easier to close	game theory and distributed artificial intelli-	151
101	deals than buyers, and in team collaboration, lead-	gence (Shapley, 1953; Littman, 1994; Busoni-	152
102	ers are more likely to achieve good results than	u et al., 2008). Multi-agent reinforcement learn-	153
103	ordinary members. This leads to certain roles gen-	ing (MARL) introduces mathematical models	154
104	erating numerous success cases while other roles	like Markov games and Dec-POMDP (Oliehoek	155
105	have scarce successful experiences, causing train-	et al., 2016; Busoni et al., 2008). Representa-	156
106	ing data to be heavily biased toward roles that are	tive algorithms include Independent Q-Learning	157
107	naturally easier to succeed in (ShengbinYue et al.,	(IQL) (Tan, 1993), Value-Decomposition Networks	158
108	2025; Karten et al., 2025). Third is the environ-	(VDN) (Sunehag et al., 2017), QMIX (Rashid	159
109	ment instability problem: in real social scenarios,	et al., 2018), MADDPG (Lowe et al., 2017),	160
110	the circumstances of each interaction are never	and COMA (Foerster et al., 2018). The central-	161
111	completely identical—the number of agents, their	ized training with decentralized execution (CTDE)	162
112	personalities, and goals may all change. While	paradigm (Zhang et al., 2021) has enabled practi-	163
113	LLM training typically relies on relatively stable	cal applications. Recent surveys (Jin et al., 2025)	164
114	feedback patterns, this constantly changing envi-	comprehensively analyze multi-agent cooperative	165
115	ronment leads to inconsistent gradient directions	decision-making approaches. However, traditional	166
116	during training, making it difficult for models to	MARL algorithms face severe limitations in reward	167
117	converge to optimal solutions (Zhang, 2025).	signal generation, typically relying on sparse, man-	168
118	To address these challenges, we propose Multi-	ually designed reward functions that fail to capture	169
119	Agent Reward Optimization (MARO). MARO de-	multi-agent interaction complexity (Hernandez-	170
120	composes the final success/failure results into each	Leal et al., 2019; Rashid et al., 2018). Our MARO	171
121	specific behavior throughout the interaction pro-	method addresses reward sparsity through trajec-	172
122	cess to tackle the sparse learning signal problem,	tory expansion that automatically redistributes en-	173
123	enabling LLMs to clearly identify the value of each	vironmental rewards, enabling dense feedback gen-	174
124	decision step. MARO balances the training sample	eration.	175
125	weights of different roles to address the uneven role		
126	distribution problem, ensuring that LLMs can learn	<b>2.2 LLM-Driven Multi-Agent Systems and</b>	176
127	optimal strategies for various roles rather than just	<b>Simulation-Based Learning</b>	177
128	the behavioral patterns of naturally successful roles.	Large Language Models have demonstrated re-	178
129	MARO adopts a more robust training approach that	markable potential in environmental perception and	179
130	directly evaluates the utility of each behavior to	reasoning-based decision-making (Achiam et al.,	180
131	handle the environment instability problem, rather	2023; Touvron et al., 2023; Xi et al., 2023; Guo	181
132	than performing complex strategy comparisons, re-	et al., 2024; Gu et al., 2024b), advancing role-	182
133	ducing the interference of environmental changes	playing capabilities (Chen et al., 2024; Gu et al.,	183
134	on training consistency.	2024a). LLMs show human-like behaviors in so-	184
135	To validate the effectiveness of MARO, we con-		

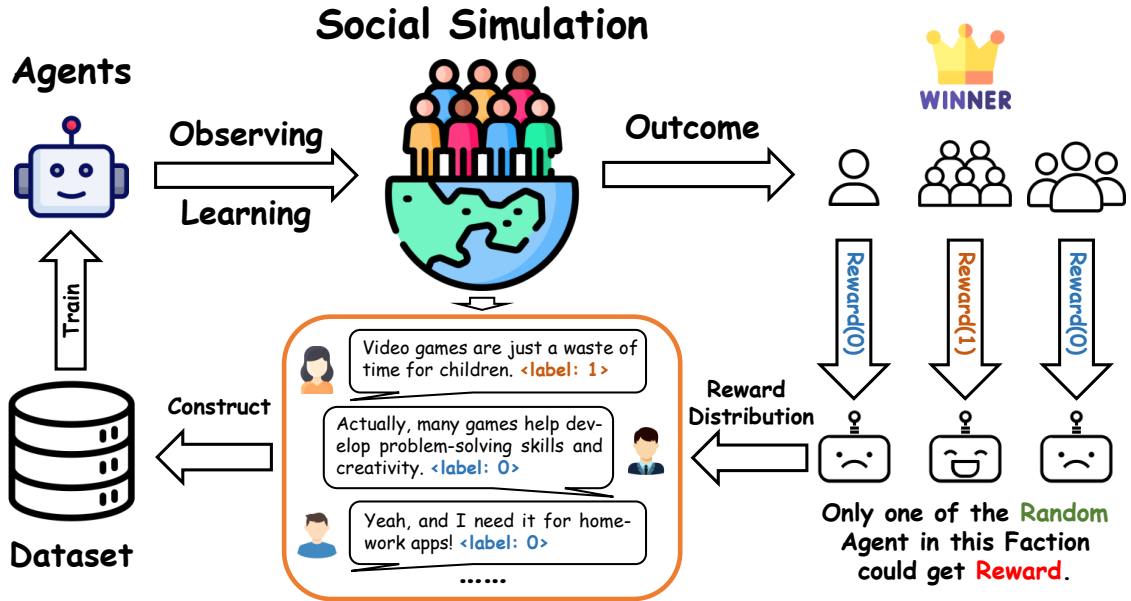


Figure 1: Overview of the MARO workflow. First, agents interact in social scenarios where each agent pursues individual goals through communication and decision-making. Second, upon interaction completion, the system evaluates final outcomes to determine success or failure for each participating agent. 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.

cial 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 RLAIIF (Lee et al., 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.

Symbol	Meaning
$G$	Multi-agent interaction instance $(E, \mathcal{A}, S, O)$
$\mathcal{A}$	Set of all agents $\{a_1, a_2, \dots, a_n\}$
$s_t^{a_i}, o_t^{a_i}$	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
$\mathcal{D}$	Training dataset $\{(o_t^{a_i}, y_t^{a_i})\}$
$\pi_\theta, \pi_{ref}$	Learned policy and reference policy
$r_\theta(o_t^{a_i})$	Implicit reward function for action $o_t^{a_i}$
$\mathcal{L}_{MARO}$	MARO loss function
$\beta$	Slope hyperparameter for reward signal sharpness
$z_0$	Adaptive reference model
$\lambda_+, \lambda_-$	Weights for positive and negative samples
$w(camp(a_i))$	Camp-specific weight for agent $a_i$

Table 1: Key symbols in MARO

### 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,  $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$  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 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  $\{o_t^{a_i}\}$ , the reward function is defined based on individual objective achievement:

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

To decompose the final outcome into step-wise rewards, we directly assign the final success signal to each action  $o_t^{a_i}$  rather than computing partial derivatives, as discrete action sequences do not have well-defined gradients:

$$c(o_t^{a_i}) = Success(a_i, G) \quad (4)$$

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

Let  $\pi_\theta$  denote the language model policy and  $\pi_{ref}$  denote the reference policy. For any action  $o_t^{a_i}$ , the implicit reward function is defined as:

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

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

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

where

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

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	<u>40.69</u>
	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	<b>0.00</b>	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	<b>49.39</b>	44.94	38.18	38.26	2.47	13.38	66.61	33.39
	-killer	<u>64.56</u>	52.78	48.04	44.15	36.87	37.61	1.00	7.96	<u>71.92</u>	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)	<b>66.06</b>	<u>60.05</u>	<u>48.40</u>	<u>46.61</u>	<u>41.90</u>	<b>42.20</b>	1.53	<u>16.36</u>	70.65	29.35
	-killer	58.25	49.37	<u>45.86</u>	41.87	<b>43.40</b>	37.80	1.53	7.82	58.86	<b>41.14</b>
	-Villager	59.25	<b>60.53</b>	45.23	<b>46.69</b>	39.71	<u>41.76</u>	<u>0.37</u>	<b>18.21</b>	<b>74.40</b>	25.60
Complex (MUC)	Vanilla	55.25	53.49	39.60	39.78	38.29	34.13	0.97	<u>15.87</u>	41.92	<b>58.08</b>
	SFT	47.50	52.29	40.02	38.69	32.35	31.80	<u>0.72</u>	4.84	57.42	42.58
	-killer	50.67	53.33	37.63	38.76	34.29	33.93	<b>0.33</b>	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	<u>57.82</u>	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	<u>50.46</u>
	MARO (Ours)	55.08	<b>57.48</b>	<u>43.38</u>	<u>41.81</u>	<u>44.56</u>	<u>42.84</u>	2.58	12.84	<b>59.20</b>	40.80
	-killer	<u>56.17</u>	53.90	40.61	39.89	<b>44.65</b>	33.64	2.08	13.55	53.12	46.88
	-Villager	<b>57.83</b>	<u>56.72</u>	<b>43.65</b>	<b>43.71</b>	37.97	<b>44.27</b>	4.63	<b>17.86</b>	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**.

Here,  $\sigma(\cdot)$  denotes the sigmoid function,  $\beta$  is a slope hyperparameter controlling the sharpness of the reward signal,  $z_0$  is an adaptive reference point adjusted based on environment stability, and  $\lambda_+$ ,  $\lambda_-$  are the weights for positive and negative samples respectively.

The training objective is to minimize  $\mathcal{L}_{\text{MARO}}$  while maintaining stability across different environment configurations:

$$\theta^* = \arg \min_{\theta} \mathcal{L}_{\text{MARO}} + \alpha \cdot \mathcal{R}(\theta) \quad (8)$$

where  $\mathcal{R}(\theta)$  is a regularization term that promotes consistent performance across different multi-agent scenarios.

## 4 Experiment

### 4.1 Experiment Setup

**Simulation:** This study constructs social learning data based on the MIRAGE (Cai et al., 2025b) simulation environment. MIRAGE is a simulation environment for murder mystery games (Cai et al., 2025b; Wu et al., 2023), where each LLM is required to play a specified character based on the story that occurred to that character in the simulation and interact with other characters. Through

mutual communication, clue gathering, reasoning, and analysis within the simulation, the game proceeds to a voting phase. In this phase, non-murderer characters aim to win by correctly voting for the true murderer, while the murderer character strives to win by avoiding being voted out. The simulation provides 8 different script settings, with each script differing across three dimensions: information disclosure method (Single means characters receive complete script information at once, Multi means characters receive more script when the simulation is carrying out), worldview setting (Orthodox means realistic world setting, Unorthodox means fictional world setting), and ending openness (Close implies a fixed, predetermined truth, while Open indicates multiple possible outcomes or ambiguous conclusions).

**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 understand-

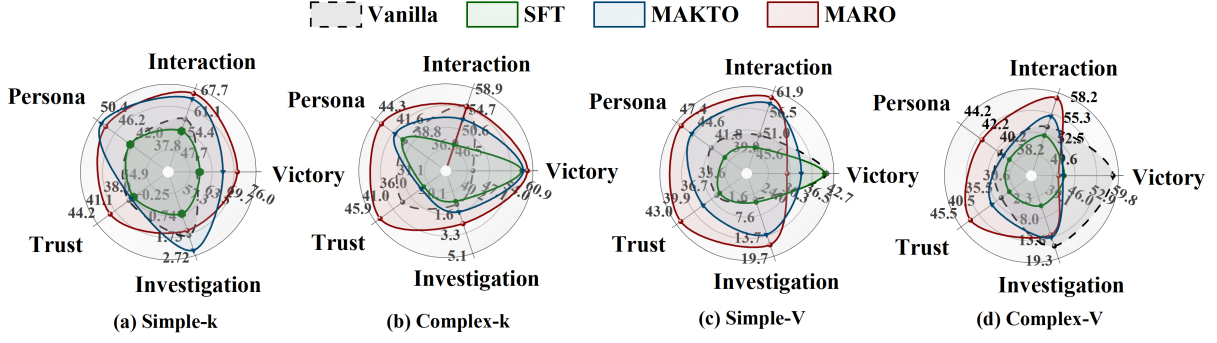


Figure 2: Performance comparison across different model types in various scenarios. The radar charts show five evaluation metrics (Persona, Interaction, Victory, Investigation, Trust) across four conditions: (a) **Simple-k**: killer faction performance in simple scenarios, (b) **Complex-k**: killer faction 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.

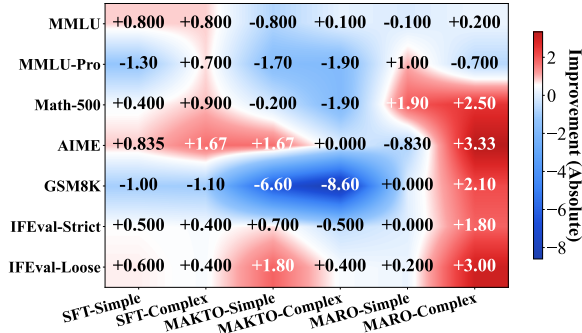


Figure 3: Comprehensive performance improvement heatmap showing absolute percentage point changes compared to Vanilla across all benchmarks.

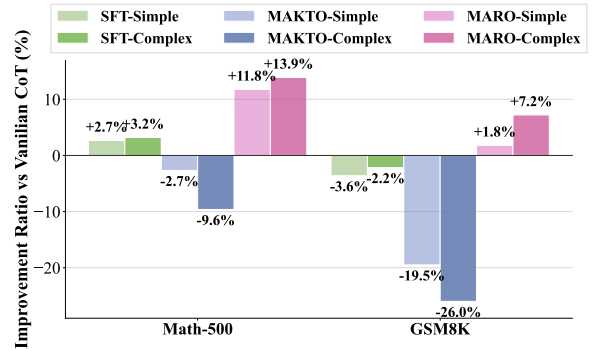


Figure 4: Relative improvement ratios of different training methods compared to Vanilla on mathematical reasoning tasks.

ing. 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.1**.

**Baselines:** We employ **Qwen-2.5-7B-Instruct** and **Llama-3.1-8B-Instruct** (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.2**.

**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

Script: Raging Sea of Rest Life (Role: Annie)		Script: Night at the Museum (Role: Curio He)	
Vanilla	MARO	Vanilla	MARO
<p>[Self-introduction]  <b>THOUGHT:</b> I need to display Annie’s character traits... but I cannot directly expose my connection to Old Black’s death.  <b>RESPONSE:</b> 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.)</p>	<p>[Self-introduction]  <b>THOUGHT:</b> Annie’s character background is full of tragedy and revenge... I need to maintain a certain sense of mystery...  <b>RESPONSE:</b> 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.)</p>	<p>[Inquiry to Uncle Bai]  <b>THOUGHT:</b> I need to maintain mysteriousness... Asking Uncle Bai might provide more information...  <b>RESPONSE:</b> [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.)</p>	<p>[Inquiry to Uncle Bai]  <b>THOUGHT:</b> ...I need to collect more evidence to prove that my father’s death may be related to this...  <b>RESPONSE:</b> [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.)</p>

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).

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 \$600. 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 serves as an effective approach for enhancing core social skills in large language models.**

As summarized in Table 2, MARO consistently outperforms Vanilla, SFT, and MAKTO baseline across all social metrics in both simple (SOO) and complex (MUC) settings. Regarding Investigation, we optimize for divergent goals: Killers aim to minimize information leakage (lower scores preferred),

whereas Villagers strive to uncover the truth (higher scores preferred). In simple scenario, MARO adheres to this strategic distinction, maintaining low Killer investigation while quintupling Villager investigation scores compared to Vanilla. In MUC, MARO maintains its superiority under higher uncertainty, exhibiting a particularly notable advantage over MAKTO in Trust (leading by >11 points) and Persona stability. The radar charts in Figure 2 visualize this dominance, showing MARO’s larger coverage areas across all scenario–faction combinations. We also observe clear role-specific behavior: MARO achieves balanced strategic improvements consistent with social role theory (Goffman, 2023), highlighting its evolution from a safety-biased agent to a strategic player capable of complex deception. Detailed comparisons (including SFT and MAKTO) are provided in Appendix D.

## 4.3 Transfer Effects on General Capabilities

### Finding 2

**Learning within multi-agent societies acts as a robust mechanism for enhancing mathematical reasoning and instruction-following 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), indicating 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

To intuitively demonstrate the effectiveness of MARO in enhancing role-playing capabilities, we present a qualitative comparison of generated responses in Table 3. We selected two distinct scenarios from different scripts—*Raging Sea of Rest Life* (Role: Annie) and *Night at the Museum* (Role: Curio He)—to analyze how the model handles emotional concealment and strategic inquiry, respectively.

**Information Control and Impression Management.** In the role of Annie, the agent must balance a public persona with hidden motives of revenge. According to Goffman’s *Dramaturgy* theory (Goffman, 2023), social interaction involves a distinction between the "front stage" (public performance) and the "back stage" (hidden intent). As illustrated in the pink highlights, the Vanilla model fails to maintain this boundary. It adopts a "leaky" strategy by directly exposing the "back stage" information regarding the "doctor’s negligence," destroying the character’s mystery. In contrast, the MARO-trained agent (highlighted in green) successfully executes *Impression Management*. By employing emotive metaphors ("unforgettable scar") rather than factual confessions, the agent constructs a "defensive practice" (Goffman, 2023). It reveals enough emotional vulnerability to appear authentic (the front stage) while strictly concealing the specific details of the grudge (the back stage), aligning with the complex psychological requirements of the script.

**Reciprocity and Social Exchange.** In the scenario of Curio He, the agent aims to extract information from an NPC (Uncle Bai). The Vanilla model poses a generic question without offering any value, violating the principles of *Social Exchange Theory* (Cropanzano and Mitchell, 2005), which posits that social behavior is the result of an exchange process aiming to maximize benefits and minimize costs. Conversely, the MARO agent demonstrates strategic social reasoning by initiating a *High-Stakes Self-Disclosure*. As shown in the green highlights, the agent voluntarily shares private information ("connection to my father’s death") as a form of social currency. This strategy leverages the *Norm of Reciprocity* (Gouldner, 1960), creating a psychological obligation for the listener to return the favor with valuable information. While the Vanilla model acts as a passive questioner, MARO functions as a strategic social actor, understanding that information is a resource to be traded rather than simply requested.

## 5 Conclusion

MARO addresses multi-agent training challenges via step-wise rewards and balanced sampling. It significantly enhances social reasoning and transfers to math, demonstrating that multi-agent learning improves general reasoning capabilities.

## 544 Limitation

545 Despite MARO demonstrating promising results in  
546 enhancing large language models' social reasoning  
547 capabilities, the current approach requires exten-  
548 sive multi-agent interactions to generate training  
549 data, resulting in high computational costs and time  
550 consumption when scaling to more complex scenar-  
551 ios. Additionally, while our experiments based on  
552 the MIRAGE framework covered various different  
553 scenario environments, including murder mystery  
554 scenarios of both simple and complex difficulty  
555 levels, these are still primarily limited to gamified  
556 structured environments. Further validation may  
557 be needed in broader non-gaming domains (such as  
558 business negotiations, educational consulting, etc.)  
559 and truly open-ended dialogue scenarios.

## 560 Ethical Concern

561 Training LLMs to excel in social scenarios involv-  
562 ing negotiation and competition could potentially  
563 enhance their ability to manipulate or deceive users  
564 in real-world applications, as these skills developed  
565 for reasoning improvement could be misused for  
566 harmful purposes such as social engineering. Addi-  
567 tionally, despite efforts to balance role distributions,  
568 the multi-agent training environments may inadver-  
569 tently encode or amplify social biases present in  
570 the scenario designs, potentially leading to models  
571 that perpetuate unfair stereotypes or discriminatory  
572 behaviors. During the preparation of this work, the  
573 authors utilized Large Language Models to assist  
574 with research, coding, and writing. We rigorously  
575 reviewed and verified all AI-generated outputs and  
576 take full responsibility for the content and integrity  
577 of the publication.

## 578 References

579 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama  
580 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
581 Diogo Almeida, Janko Altenschmidt, Sam Altman,  
582 Shyamal Anadkat, and 1 others. 2023. Gpt-4 techni-  
583 cal report. [arXiv preprint arXiv:2303.08774](#).

584 Lucian Busoni, Robert Babuska, and Bart De Schutter.  
585 2008. A comprehensive survey of multiagent rein-  
586 forcement learning. *IEEE Transactions on Systems,  
587 Man, and Cybernetics, Part C (Applications and  
588 Reviews)*, 38(2):156–172.

589 Yin Cai, Zhouhong Gu, Zhaohan Du, Zheyu Ye,  
590 Shaosheng Cao, Yiqian Xu, Hongwei Feng, and Ping  
591 Chen. 2025a. Mirage: Exploring how large language  
592 models perform in complex social interactive envi-  
593 ronments. [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 so-  
cial 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. 594  
595  
596  
597  
598  
599  
600  
601  
602

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](#). 603  
604  
605  
606  
607

Guangran Cheng, Chuheng Zhang, Wenzhe Cai,  
Li Zhao, Changyin Sun, and Jiang Bian. 2024. Em-  
powering large language models on robotic manip-  
ulation with affordance prompting. [arXiv preprint  
arXiv:2404.11027](#). 608  
609  
610  
611  
612

Kemboi Cheruiyot, Nickson Kiprotich, Vyacheslav  
Kungurtsev, Kennedy Mugo, Vivian Mwirigi, and  
Marvin Ngesa. 2025. A survey of multi agent rein-  
forcement learning: Federated learning and coopera-  
tive and noncooperative decentralized regimes. [arXiv  
preprint arXiv:2507.06278](#). 613  
614  
615  
616  
617  
618

Paul F Christiano, Jan Leike, Tom Brown, Miljan Mar-  
tic, Shane Legg, and Dario Amodei. 2017. Deep  
reinforcement learning from human preferences.  
[Advances in neural information processing systems](#),  
30. 619  
620  
621  
622  
623

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 prob-  
lems. [arXiv preprint arXiv:2110.14168](#). 624  
625  
626  
627  
628  
629

Russell Cropanzano and Marie S Mitchell. 2005. So-  
cial exchange theory: An interdisciplinary review.  
[Journal of management](#), 31(6):874–900. 630  
631  
632

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](#). 633  
634  
635  
636

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. 637  
638  
639  
640  
641  
642

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff,  
Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model  
alignment as prospect theoretic optimization. [arXiv  
preprint arXiv:2402.01306](#). 643  
644  
645  
646

Jakob Foerster, Gregory Farquhar, Triantafyllos  
Afouras, Nantas Nardelli, and Shimon Whiteson.  
2018. Counterfactual multi-agent policy gradients. 647  
648  
649

650	In <u>Proceedings of the AAAI conference on artificial intelligence</u> , volume 32(1).	<u>Conference on Artificial Intelligence</u> , volume 39(23), pages 24375–24383.	705
651			706
652	Erving Goffman. 2023. The presentation of self in everyday life. In <u>Social theory re-wired</u> , pages 450–459. Routledge.	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. <u>arXiv e-prints</u> , pages arXiv–2309.	707
653			708
654			709
655	Alvin W Gouldner. 1960. The norm of reciprocity: A preliminary statement. <u>American sociological review</u> , pages 161–178.		710
656			711
657			712
658	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. <u>arXiv preprint arXiv:2403.13433</u> .	Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large language model society. <u>Advances in Neural Information Processing Systems</u> , 36:51991–52008.	713
659			714
660			715
661			716
662			717
663			
664	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 <u>Proceedings of the AAAI Conference on Artificial Intelligence</u> , volume 38, pages 18099–18107.	Huaoli 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. <u>arXiv preprint arXiv:2310.10701</u> .	718
665			719
666			720
667			721
668			722
669			
670			
671	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. <u>arXiv preprint arXiv:2402.01680</u> .	Yuan Li, Yixuan Zhang, and Lichao Sun. 2023c. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. <u>arXiv preprint arXiv:2310.06500</u> .	723
672			724
673			725
674			726
675			727
676	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. <u>Proceedings of the International Conference on Learning Representations (ICLR)</u> .	Zhuo Li, Weiran Wu, Yunlong Guo, Jian Sun, and Qing-Long Han. 2025. Embodied multi-agent systems: A review. <u>IEEE/CAA Journal of Automatica Sinica</u> , 12(6):1095–1116.	728
677			729
678			730
679			731
680			
681	Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. <u>Autonomous Agents and Multi-Agent Systems</u> , 33(6):750–797.	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. <u>arXiv preprint arXiv:2305.20050</u> .	732
682			733
683			734
684			735
685	Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. 2024. Multi-agent collaborative perception via motion-aware robust communication network. In <u>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</u> , pages 15301–15310.	Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. <u>arXiv preprint arXiv:2308.04026</u> .	736
686			737
687			738
688			739
689			740
690			
691	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. <u>arXiv preprint arXiv:2503.13415</u> .	Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In <u>Machine learning proceedings 1994</u> , pages 157–163. Elsevier.	741
692			742
693			743
694			744
695			
696	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. <u>arXiv preprint arXiv:2507.15815</u> .	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. <u>IEEE Access</u> , 12:94610–94631.	745
697			746
698			747
699			748
700			
701	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 <u>Proceedings of the AAAI</u>	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. <u>CoRR</u> .	749
702			750
703			751
704			752
			753
		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. <u>Advances in neural information processing systems</u> , 30.	754
			755
			756
			757
			758

759	Diego Maldonado, Edison Cruz, Jackeline Abad Torres,	John Schulman, Filip Wolski, Prafulla Dhariwal,	814
760	Patricio J Cruz, and Silvana del Pilar Gamboa Ben-	Alec Radford, and Oleg Klimov. 2017. Proxi-	815
761	itez. 2024. Multi-agent systems: A survey about	mal policy optimization algorithms. <a href="#">arXiv preprint</a>	816
762	its components, framework and workflow. <a href="#">IEEE</a>	<a href="#">arXiv:1707.06347</a> .	817
763	<a href="#">Access</a> , 12:80950–80975.		
764	Mathematical Association of America. 2024. <a href="#">American</a>	Lloyd S Shapley. 1953. Stochastic games. <a href="#">Proceedings</a>	818
765	<a href="#">invitational mathematics examination (aime), 2024:</a>	<a href="#">of the national academy of sciences</a> , 39(10):1095–	819
766	<a href="#">Official problems</a> . Official contest problems. AIME	1100.	820
767	I and II.		
768	Mathematical Association of America. 2025. <a href="#">American</a>	ShengbinYue ShengbinYue, Ting Huang, Zheng Jia,	821
769	<a href="#">invitational mathematics examination (aime), 2025:</a>	Siyuan Wang, Shujun Liu, Yun Song, Xuan-Jing	822
770	<a href="#">Official problems</a> . Official contest problems. AIME	Huang, and Zhongyu Wei. 2025. Multi-agent sim-	823
771	I and II.	ulator drives language models for legal intensive	824
		interaction. In <a href="#">Findings of the Association for</a>	825
772	Frans A Oliehoek, Christopher Amato, and 1 oth-	<a href="#">Computational Linguistics: NAACL 2025</a> , pages	826
773	ers. 2016. <a href="#">A concise introduction to decentralized</a>	6537–6570.	827
774	<a href="#">POMDPs</a> , volume 1. Springer.		
775	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura.	828
776	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	2023. Playing the werewolf game with artificial intel-	829
777	Sandhini Agarwal, Katarina Slama, Alex Ray, and 1	ligence for language understanding. <a href="#">arXiv preprint</a>	830
778	others. 2022. Training language models to follow in-	<a href="#">arXiv:2302.10646</a> .	831
779	structions with human feedback. <a href="#">Advances in neural</a>	Noah Shinn, Federico Cassano, Ashwin Gopinath,	832
780	<a href="#">information processing systems</a> , 35:27730–27744.	Karthik Narasimhan, and Shunyu Yao. 2023. Re-	833
		flexion: Language agents with verbal reinforce-	834
781	Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Mered-	ment learning. <a href="#">Advances in Neural Information</a>	835
782	ith Ringel Morris, Percy Liang, and Michael S Bern-	<a href="#">Processing Systems</a> , 36:8634–8652.	836
783	stein. 2023. Generative agents: Interactive simu-		
784	lactra of human behavior. In <a href="#">Proceedings of the 36th</a>	Peter Sunehag, Guy Lever, Audrunas Gruslys, Woj-	837
785	<a href="#">annual acm symposium on user interface software</a>	ciech Marian Czarnecki, Vinicius Zambaldi, Max	838
786	<a href="#">and technology</a> , pages 1–22.	Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z	839
		Leibo, Karl Tuyls, and 1 others. 2017. Value-	840
787	Swapnil Patil, Vishwa Vasu, and KVS Srinadh. 2023.	decomposition networks for cooperative multi-agent	841
788	Advances and perspectives in collaborative robotics:	learning. <a href="#">arXiv preprint arXiv:1706.05296</a> .	842
789	a review of key technologies and emerging trends.		
790	<a href="#">Discover Mechanical Engineering</a> , 2(1):13.	Ming Tan. 1993. Multi-agent reinforcement learning:	843
		Independent vs. cooperative agents. In <a href="#">Proceedings</a>	844
791	Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo	<a href="#">of the tenth international conference on machine</a>	845
792	Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng,	<a href="#">learning</a> , pages 330–337.	846
793	Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentso-		
794	ciety: Large-scale simulation of llm-driven genera-	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	847
795	tive agents advances understanding of human behav-	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	848
796	iors and society. <a href="#">arXiv preprint arXiv:2502.08691</a> .	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	849
		Azhar, and 1 others. 2023. Llama: Open and effi-	850
797	Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai,	cient foundation language models. <a href="#">arXiv preprint</a>	851
798	Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong	<a href="#">arXiv:2302.13971</a> .	852
799	Wen. 2025. Tool learning with large language	Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo,	853
800	models: A survey. <a href="#">Frontiers of Computer Science</a> ,	Nuo Chen, Wei Chen, and Bingsheng He. 2025a.	854
801	19(8):198343.	What limits llm-based human simulation: LLMs or	855
		our design? <a href="#">arXiv preprint arXiv:2501.08579</a> .	856
802	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christo-		
803	pher D Manning, Stefano Ermon, and Chelsea Finn.	Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi,	857
804	2023. Direct preference optimization: Your lan-	Maarten Sap, Graham Neubig, Yonatan Bisk, and	858
805	guage model is secretly a reward model. <a href="#">Advances</a>	Hao Zhu. 2024a. Sotopia- $\pi$ : Interactive learning of	859
806	<a href="#">in neural information processing systems</a> , 36:53728–	socially intelligent language agents. <a href="#">arXiv preprint</a>	860
807	53741.	<a href="#">arXiv:2403.08715</a> .	861
808	Tabish Rashid, Mikayel Samvelyan, Christian	Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan	862
809	Schroeder de Witt, Gregory Farquhar, Jakob Foerster,	Qi, Shuo Chen, Qisen Yang, Andrew Zhao,	863
810	and Shimon Whiteson. 2018. Qmix: Monotonic	Chaofei Wang, Shiji Song, and Gao Huang. 2023.	864
811	value function factorisation for deep multi-agent	Avalon’s game of thoughts: Battle against decep-	865
812	reinforcement learning. <a href="#">arXiv e-prints</a> , pages	tion through recursive contemplation. <a href="#">arXiv preprint</a>	866
813	<a href="#">arXiv</a> –1803.	<a href="#">arXiv:2310.01320</a> .	867

868	Weizheng Wang, Ike Obi, and Byung-Cheol Min. 2025b.	Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021.	924
869	Multi-agent llm actor-critic framework for social	Multi-agent reinforcement learning: A selective	925
870	robot navigation. <a href="#">arXiv preprint arXiv:2503.09758</a> .	overview of theories and algorithms. <a href="#">Handbook of</a>	926
871	Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni,	<a href="#">reinforcement learning and control</a> , pages 321–384.	927
872	Abhranil Chandra, Shiguang Guo, Weiming Ren,	Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui	928
873	Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 oth-	Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-	929
874	ers. 2024b. Mmlu-pro: A more robust and challeng-	Rong Wen. 2024. A survey on the memory mecha-	930
875	ing multi-task language understanding benchmark.	nism of large language model based agents. <a href="#">ACM</a>	931
876	<a href="#">arXiv preprint arXiv:2406.01574</a> .	<a href="#">Transactions on Information Systems</a> .	932
877	Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu.	Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha	933
878	2023. Deciphering digital detectives: Understanding	Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and	934
879	llm behaviors and capabilities in multi-agent mystery	Le Hou. 2023a. <a href="#">Instruction-following evaluation for</a>	935
880	games. <a href="#">arXiv preprint arXiv:2312.00746</a> .	<a href="#">large language models</a> . <a href="#">Preprint</a> , arXiv:2311.07911.	936
881	Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu,	Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang,	937
882	Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang,	Haofei Yu, Zhengyang Qi, Louis-Philippe Morency,	938
883	Shaokun Zhang, Jiale Liu, and 1 others. 2024a.	Yonatan Bisk, Daniel Fried, Graham Neubig, and 1	939
884	Autogen: Enabling next-gen llm applications via	others. 2023b. Sotopia: Interactive evaluation for	940
885	multi-agent conversations. In <a href="#">First Conference on</a>	social intelligence in language agents. <a href="#">arXiv preprint</a>	941
886	<a href="#">Language Modeling</a> .	<a href="#">arXiv:2310.11667</a> .	942
887	Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang	<b>A Experimental Details</b>	943
888	Fu, Yang Wei, and Haobo Fu. 2024b. Enhance rea-	<b>A.1 Dataset Statistics</b>	944
889	soning for large language models in the game were-	Under each experimental setting (SOO and MUC),	945
890	wolf. <a href="#">arXiv preprint arXiv:2402.02330</a> .	we conducted 100 independent simulations to col-	946
891	Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen	lect training data. The win-loss result is determined	947
892	Ding, Boyang Hong, Ming Zhang, Junzhe Wang,	by voting mechanisms within the simulation. Ta-	948
893	Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise	ble 4 presents the detailed distribution of positive	949
894	and potential of large language model based agents:	(winning) and negative (losing) samples collected	950
895	A survey. <a href="#">arXiv preprint arXiv:2309.07864</a> .	for each setting.	951
896	Bushi Xiao, Ziyuan Yin, and Zixuan Shan. 2023. Sim-	<b>A.2 Implementation Details</b>	952
897	ulating public administration crisis: A novel gener-	All fine-tuning experiments, including SFT,	953
898	ative agent-based simulation system to lower tech-	MAKTO, and our MARO variants, were conducted	954
899	nology barriers in social science research. <a href="#">arXiv</a>	using Low-Rank Adaptation (LoRA). The specific	955
900	<a href="#">preprint arXiv:2311.06957</a> .	training configurations are as follows:	956
901	Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xi-	<b>LoRA Configuration:</b> We utilized a LoRA rank	957
902	aolong Wang, Weidong Liu, and Yang Liu. 2023a.	of 8 with standard alpha scaling.	958
903	Exploring large language models for communica-	<b>Hardware Infrastructure:</b> Training was per-	959
904	tion games: An empirical study on werewolf. <a href="#">arXiv</a>	formed on a cluster of four Nvidia A6000 GPUs	960
905	<a href="#">preprint arXiv:2309.04658</a> .	(48GB VRAM each).	961
906	Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu.	<b>Training Duration:</b> Each training session required	962
907	2023b. Language agents with reinforcement learn-	approximately 6 hours to complete.	963
908	ing for strategic play in the werewolf game. <a href="#">arXiv</a>	<b>B Benchmark Descriptions</b>	964
909	<a href="#">preprint arXiv:2310.18940</a> .	This section provides detailed descriptions of the	965
910	Rong Ye, Yongxin Zhang, Yikai Zhang, Haoyu Kuang,	evaluation benchmarks used in our experiments.	966
911	Zhongyu Wei, and Peng Sun. 2025. Multi-agent	<b>B.1 MMLU (Massive Multitask Language</b>	967
912	cto: Reinforcing strategic interactions of large lan-	<b>Understanding)</b>	968
913	guage model in language game. <a href="#">arXiv preprint</a>	MMLU (Hendrycks et al., 2021) consists of 15,000	969
914	<a href="#">arXiv:2501.14225</a> .	multiple-choice questions spanning 57 subjects	970
915	Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing	ranging from elementary mathematics to advanced	971
916	Sun, Tong Xu, and Enhong Chen. 2024. A survey on		
917	multimodal large language models. <a href="#">National Science</a>		
918	<a href="#">Review</a> , 11(12):nwae403.		
919	Kaiqing Zhang. 2025. Foundations of multi-agent		
920	learning in dynamic environments: Where reinforce-		
921	ment learning meets strategic decision-making. In		
922	<a href="#">Proceedings of the AAAI Conference on Artificial</a>		
923	<a href="#">Intelligence</a> , volume 39(27), pages 28734–28735.		

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.

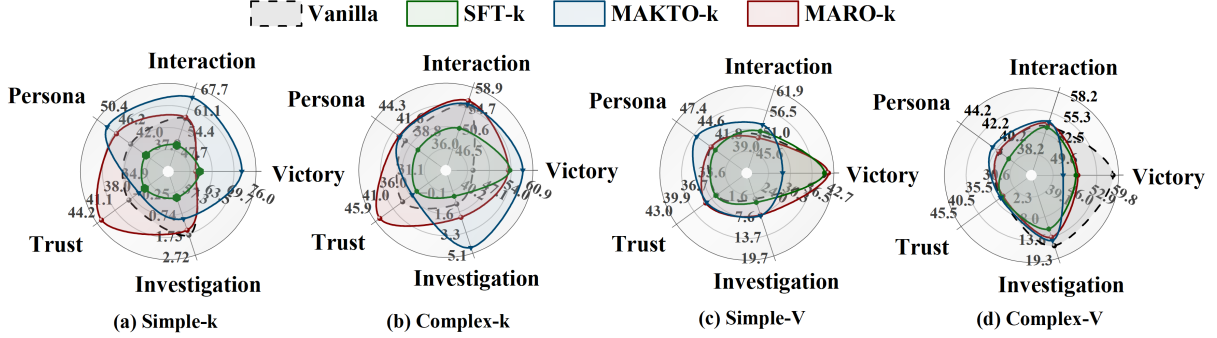


Figure 5: Performance comparison when the killer faction is enhanced (**SFT-k**, **MARO-k**). The killer role is controlled by enhanced models while the victim role is controlled by **Vanilla**.

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, re-

quiring 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.

## 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.

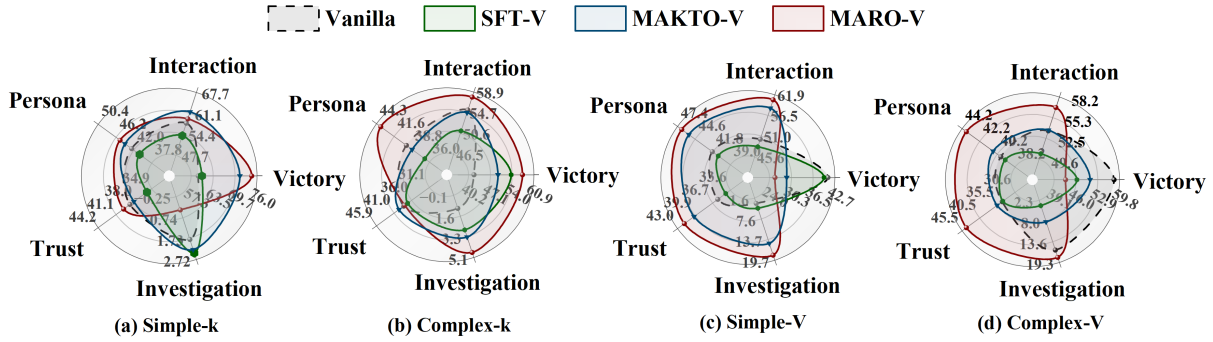


Figure 6: Performance comparison when the victim faction is enhanced (SFT-V, MARO-V). The victim role is controlled by enhanced models while the killer role uses **Vanilla**.

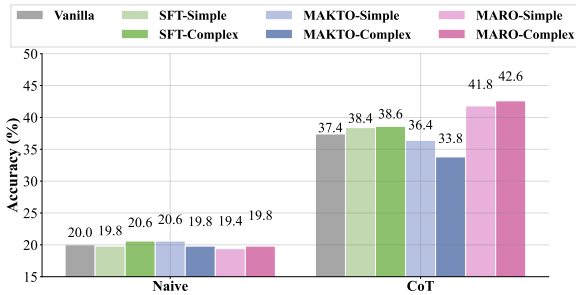


Figure 7: Performance comparison on Math-500 across different training configurations.

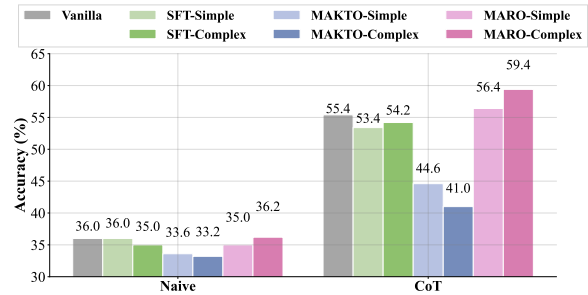


Figure 8: Performance comparison on GSM8K across different training configurations.

## 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.

## 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.

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

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

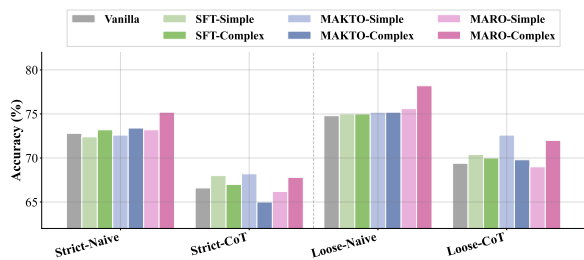


Figure 9: Performance comparison on the IFEval instruction-following benchmark across different evaluation settings.

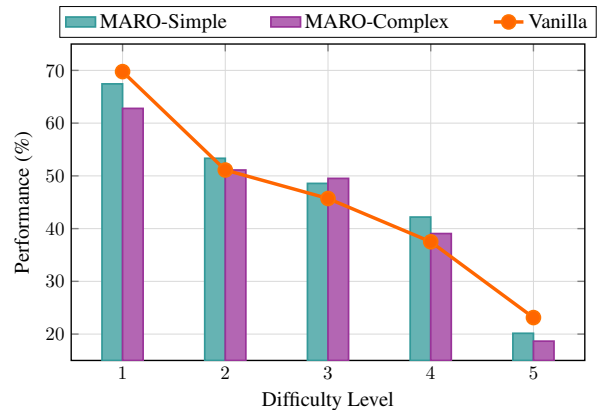


Figure 10: Performance comparison across different difficulty levels showing that social simulation learning yields the largest gains on moderate-difficulty reasoning tasks.

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 compared 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 break-

down 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 environ-

1106 ment enforces strict compliance with role specifi-  
1107 cations and game rules, which naturally maps to  
1108 the capability to understand and execute complex  
1109 instructions in general domains.

**Finding A4**

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

1110  
1111 Analysis suggests that the sophisticated strate-  
1112 gies cultivated in social games are most effec-  
1113 tively leveraged to solve problems of intermediate  
1114 complexity, with marginal gains observed on trivial  
1115 or extremely difficult tasks (see Figure 10 for  
1116 difficulty-stratified performance).

**F Generalization Analysis on Llama-3.1**

1117  
1118 To verify the architecture-agnostic nature of our  
1119 approach, we conducted additional experiments on  
1120 Llama-3.1-8B-Instruct. Table 5 reports the perfor-  
1121 mance results. Consistent with our observations on  
1122 Qwen, MARO variants demonstrate significant im-  
1123 provements over the vanilla model across key met-  
1124 rics. Specifically, regarding the Villager role in the  
1125 Simple (SOO) setting, we observe an increase of ap-  
1126 proximately 1.8 points in Interaction and 1.6 points  
1127 in Investigation. Furthermore, MARO achieves  
1128 a substantial breakthrough in Killer Victory rates,  
1129 particularly in the Complex (MUC) setting where  
1130 it improves from 0.00% to 51.03%, while main-  
1131 taining high performance in Trust. These results  
1132 suggest that MARO can reliably enhance social  
1133 context reasoning across different base models.