MARO: Let Large Language Models Learn in Social Scenarios

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

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 conducted extensive experiments in simulated social environments. The results demonstrate that MARO significantly enhances large language models' 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, complex social environments proved more effective than simple ones in promoting general reasoning capability enhancement.

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 so-

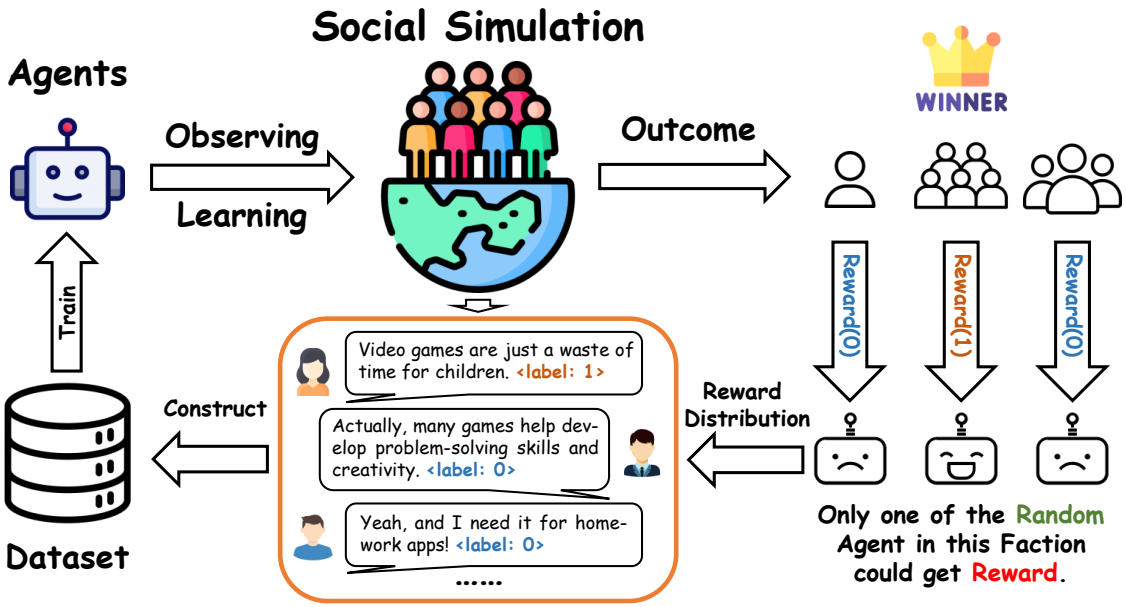

Figure 1: Overview of the MARO workflow. First, agents interact in social scenarios where each agent pursues individual goals through communication and decision-making. Second, upon interaction completion, the system evaluates final outcomes to determine success or failure for each participating agent. 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 RLAIF (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 $\{a_1, a_2, \ldots, a_n\}$
$s_t^{a_i}, o_t^{a_i}$	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 $\{(o_t^{a_i}, y_t^{a_i})\}$
$\pi_\theta, \pi_{\text{ref}}$	Learned policy and reference policy
$r_\theta(o_t^{a_i})$	Implicit reward function for action $o_t^{a_i}$
$\mathcal{L}_{\text{MARO}}$	MARO loss function
β	Slope hyperparameter for reward signal sharpness
z_0	Adaptive reference model
λ_+, λ_-	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, $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ 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 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 $Success(a_i, G)$:

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

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:

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

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:

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

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:

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

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

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

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

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

Script Type	Method	Interaction	Persona	Trust	Investigation	Victory
Simple (SOO)	**Vanilla**	51.29	42.40	36.32	4.85	40.69
	SFT	48.01	40.09	35.78	4.67	40.00
	MARO	**61.54**	**46.99**	**42.21**	17.89	29.35
	SFT-k	50.82	40.74	36.03	4.10	40.00
	MARO-k	50.80	42.48	38.76	9.35	**41.14**
	SFT-V	49.13	40.82	34.72	6.93	39.63
	MARO-V	60.48	46.46	41.56	**18.58**	25.60
Complex (MUC)	**Vanilla**	53.88	39.81	36.24	16.84	**58.08**
	SFT	51.85	38.89	32.65	5.56	42.58
	MARO	**57.34**	41.90	**43.93**	15.41	40.80
	SFT-k	53.13	38.67	35.25	11.62	46.33
	MARO-k	54.30	39.96	36.52	15.64	46.88
	SFT-V	50.55	39.06	34.48	6.48	46.30
	MARO-V	56.80	**43.70**	43.74	**22.50**	42.68

Table 2: Main Social Capability results of **Vanilla**, **SFT**, and **MARO**. The table presents evaluation scores (in %) across five different metrics. **Bolded values** indicate the best result for each metric, while underlined values denote the second-best performance. **Vanilla**, **SFT** and **MARO** represent settings where both roles are played by **Vanilla**, **SFT** or **MARO**, respectively. **SFT-k** and **MARO-k** indicate settings where the killer is controlled by **SFT** or **MARO** and the Victim is controlled by **Vanilla**. Conversely, **SFT-V** and **MARO-V** denote cases where the Victim is controlled by **SFT** or **MARO** and the killer is controlled by **Vanilla**.

Here, $\sigma(\cdot)$ denotes the sigmoid function, β is a slope hyperparameter controlling the sharpness of the reward signal, z_0 is an adaptive reference point adjusted based on environment stability, and λ_+, λ_- 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:

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

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 in the simulation, they aim to identify the real murderer. Non-murderer characters need to win by voting to identify the murderer, while the murderer character needs 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 script information in segments), worldview setting (Orthodox means realistic world setting, Unorthodox means fictional world setting), and ending openness (Open means open ending, Close means closed ending).

Training: We selected two representative scripts for experiments: Single-Orthodox-Open (SOO) and Multi-Unorthodox-Close (MUC). SOO adopts one-time information disclosure and realistic world setting, reducing the complexity of information processing and world understanding, so we call the model trained on it MARO-Simple. MUC combines progressive information disclosure, fictional world setting, and closed ending, requiring the model to have stronger dynamic adaptation and abstract reasoning capabilities, so we call the model trained on it MARO-Complex. Under each script setting, we conducted 100 independent simulations, with each character able to speak up to five times in turn during each round of interaction. After each

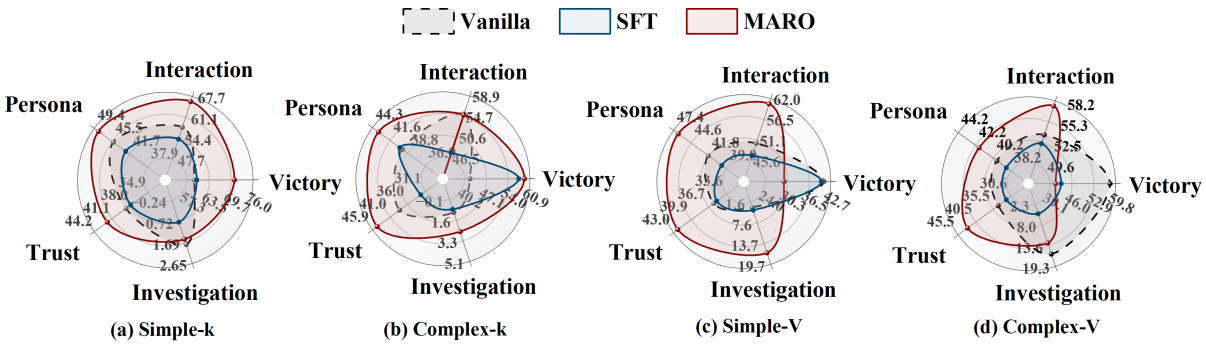

Figure 2: Performance comparison across different model types in various scenarios. The radar charts show five evaluation metrics (Persona, Interaction, Victory, Investigation, Trust) across four conditions: (a) **Simple-k**: killer faction performance in simple scenarios, (b) **Complex-k**: killer faction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

simulation ends, the final win-loss result is determined by voting, with the winning side's behaviors labeled as positive samples and the losing side's behaviors labeled as negative samples. Finally, we collected 1,204 positive samples and 1,586 negative samples under the SOO setting, and 1,043 positive samples and 3,149 negative samples under the MUC setting. We use the MARO method to train the baseline model Qwen-2.5-7B-Instruct respectively, using LoRA configuration with rank 8 on four Nvidia A6000 48G GPUs, with each model training for approximately 6 hours. Meanwhile, we remove negative samples from the dataset for supervised fine-tuning (SFT) training as a comparative baseline to validate the effectiveness of MARO.

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 $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.

For general reasoning capability evaluation, we adopt six benchmark tests to evaluate models' capabilities in factual memory, mathematical reasoning, and instruction following: **MMLU** (Hendrycks et al., 2021) contains 15,000 multiple-choice questions across 57 subjects for cross-disciplinary knowledge evaluation. **MMLU-pro** (Wang et al., 2024b) provides over 10,000 challenging questions with difficult distractors. **MATH-500** (Lightman et al., 2023) includes 500 high-difficulty competition mathematics problems. **GSM8K** (Cobbe et al., 2021) covers 8,800 elementary-level arithmetic word problems. **AIME** (of America, 2024, 2025) includes 60 advanced mathematical competition problems from AIME-2024 and AIME-2025. **IFEval** (Zhou et al., 2023a) provides over 1,000 instruction-following tasks with automatic scoring. We simultaneously compare model performance on these benchmarks under both naive methods and chain-of-thought (CoT) (Wei et al., 2022) guidance.

4.2 Social Interaction Performance Analysis

As shown in Table 2, MARO achieved superior scores: Interaction 61.54 (vs. Vanilla's 51.29 and SFT's 48.01), Persona 46.99 (vs. Vanilla's 42.40 and SFT's 40.09), and Trust 42.21 (vs. Vanilla's 36.32 and SFT's 35.78). In Complex scenarios,

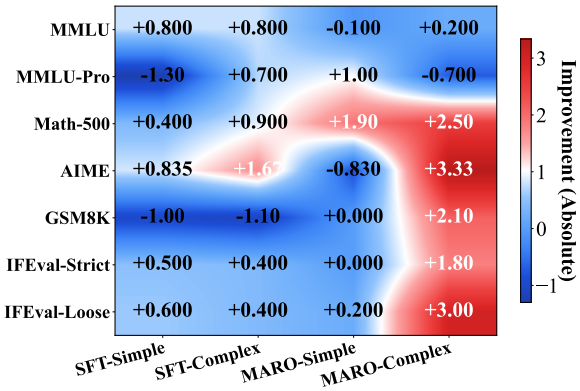

Figure 3: Comprehensive performance improvement heatmap showing absolute percentage point changes compared to Vanilla across all benchmarks.

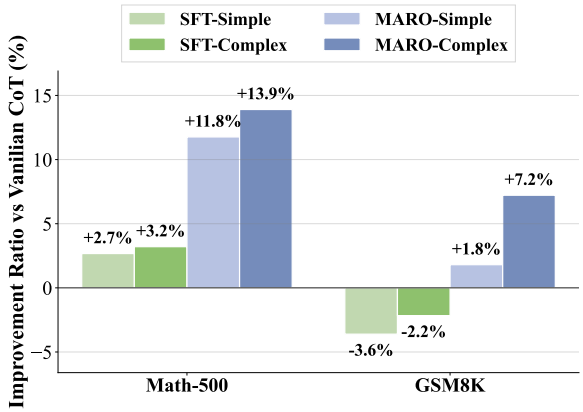

Figure 4: Relative improvement ratios of different training methods compared to Vanilla on mathematical reasoning tasks.

MARO similarly excelled with scores of 57.34, 41.90, and 43.93 respectively. The radar charts in Figure 2 provide a more intuitive visualization, where MARO (red area) demonstrates larger coverage areas across all four scenario-faction combinations, particularly expanding significantly in the Interaction, Persona, and Trust dimensions. This improvement likely stems from MARO's ability to learn complex strategies from real social scenarios through multi-agent interactions, including how to adjust communication styles based on dialogue history, maintain role consistency, and make trust judgments in uncertain environments.

Finding 1

MARO significantly enhances AI agents' social reasoning capabilities through training in multi-agent social environments, including core social skills such as understanding others' behaviors, identifying personality traits, and building trust relationships.

Table 2 demonstrates MARO's consistent advantages across both scenario complexities and multiple social reasoning dimensions. In Simple scenarios, MARO achieved superior scores across four key metrics: Interaction (61.54 vs. Vanilla's 51.29), Persona (46.99 vs. 42.40), Trust (42.21 vs. 36.32), and Investigation (17.89 vs. 4.85). In Complex scenarios, MARO maintained its advantages with scores of 57.34, 41.90, 43.93, and 15.41 respectively, demonstrating robust performance across varying environmental uncertainties.

The improvement in Investigation capability was particularly remarkable, representing a qualitative leap in agents' complex problem-solving abilities. As shown in Figure 10 and Figure 9, the optimal situational reasoning capability was achieved by the **MARO-V** configuration (Simple scenario 18.58, Complex scenario 22.50), with the Victim faction's MARO training showing especially prominent Investigation improvements.

Finding 2

MARO demonstrates comprehensive performance improvements across diverse social reasoning tasks and scenario complexities, with particularly notable enhancements in situational investigation capabilities.

The radar charts in Figure 2 provide intuitive visualization of MARO's comprehensive improvements, where MARO (red area) demonstrates larger coverage areas across all four scenario-faction combinations. This systematic enhancement likely stems from MARO's ability to learn complex social strategies through multi-agent interactions, including adaptive communication styles, role consistency maintenance, and trust judgments under uncertainty.

However, MARO's Victory shows clear role-specific differentiation: significantly improved in killer scenarios while declining in Victim scenarios. This phenomenon aligns with social role theory (Goffman, 2023)—MARO adapts behavioral strategies according to social roles, adopting competitive strategies in killer roles while favoring cooperative prosocial behaviors in Victim roles.

For detailed analysis of Interaction capabilities comparisons, it can be referred to Appendix B.

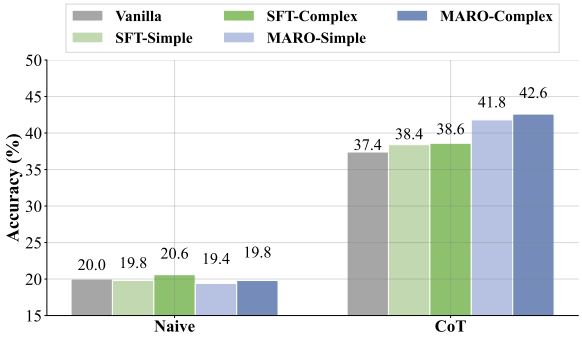

Figure 5: Performance comparison on Math-500 across different training configurations.

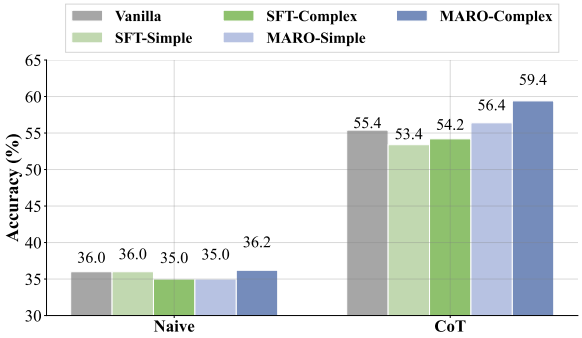

Figure 6: Performance comparison on GSM8K across different training configurations.

4.3 Transfer Effects on General Capabilities

As shown in Figure 3 reveals differentiated improvement effects: models achieve the most significant benefits on mathematics-related tasks (2.5-3.33 percentage points) and instruction following tasks (3.0 percentage points), while showing limited improvements on knowledge-intensive tasks like MMLU. This occurs possibly because multi-agent games primarily exercise advanced cognitive skills such as dynamic reasoning, strategic planning, and constraint compliance, which are highly consistent with mathematical problem-solving and instruction following but provide limited benefit for knowledge memorization tasks.

Finding 3

Learning in Social simulation substantially improves mathematical reasoning and instruction following capabilities of large language models.

Figures 5 and 6 show consistent patterns where MARO-Complex outperforms MARO-Simple across both benchmarks. On Math-500, MARO-Complex achieves 59.4% accuracy with Chain-of-Thought prompting versus MARO-Simple's 56.4%, while on GSM8K, MARO-Complex reaches 42.6% compared to 41.8%. Both configurations substantially exceed their SFT counterparts and vanilla baseline, with MARO-Complex showing the most pronounced improvements in Chain-of-Thought scenarios.

Figure 4 confirms complex environments' superiority, with MARO-Complex achieving 13.9% relative improvement on Math-500 versus MARO-Simple's 11.8%, and more pronounced differences on GSM8K (7.2% vs 1.8%). This consistent pattern confirms that environmental complexity directly correlates with cognitive skill transfer magnitude to quantitative domains.

Finding 4

Rich social environments demonstrate superior effectiveness over minimal environments in enhancing mathematical reasoning capabilities, with complex multi-agent interactions facilitating more robust cognitive skill transfer to quantitative problem-solving tasks.

Complex social configurations excel due to the Multi-Unorthodox-Close setting's cognitive demands. These environments require sophisticated reasoning under progressive information disclosure, fictional world modeling, and closed-ending resolution, exercising higher-order functions including working memory management and multi-step logical inference. These patterns align closely with mathematical reasoning's structured problem-solving processes, creating stronger cognitive bridges between social interaction learning and quantitative capabilities.

Additional detailed experimental results and analyses for general capabilities are provided in Appendix C.

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' social reasoning capabilities, the current approach requires extensive multi-agent interactions to generate training data, resulting in high computational costs and time consumption when scaling to more complex scenarios. Additionally, while our experiments based on the MIRAGE framework covered various different scenario environments, including murder mystery scenarios of both simple and complex difficulty levels, these are still primarily limited to gamified structured environments. Further validation may be needed in broader non-gaming domains (such as business negotiations, educational consulting, etc.) and truly open-ended dialogue scenarios.

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.

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.

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.

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.

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.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

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 Benchmark Descriptions

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

A.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.

A.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 greater 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.

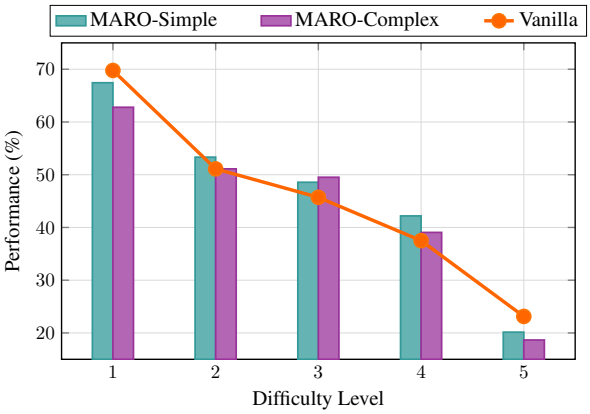

Figure 7: Performance comparison across different difficulty levels showing MARO's optimal performance on moderate-difficulty reasoning tasks.

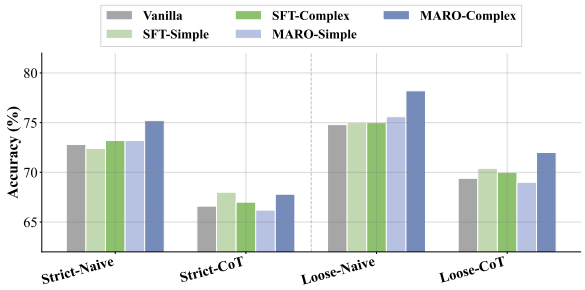

Figure 8: Performance comparison on IFEval instruction following benchmark across different evaluation settings.

A.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.

A.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 to 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.

A.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.

A.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.

B Detailed Analysis of Social Interaction

Finding A1

Supervised Fine-Tuning (SFT) shows limited effectiveness for comprehensive social capability enhancement in our multi-agent social reasoning scenarios, with performance degradation observed in several core interaction metrics.

Within the context of our SOTOPIA-based evaluation framework, SFT demonstrated suboptimal performance compared to the vanilla baseline. In Simple scenarios, SFT showed declines in Interaction (48.01 vs 51.29), Persona (40.09 vs 42.40), and Trust (35.78 vs 36.32) capabilities. Similar patterns emerged in Complex scenarios, where multiple metrics fell below the Vanilla baseline.

The radar chart comparisons in Figure 2 illustrate this phenomenon more intuitively: across all subfigures, SFT (blue area) consistently shows

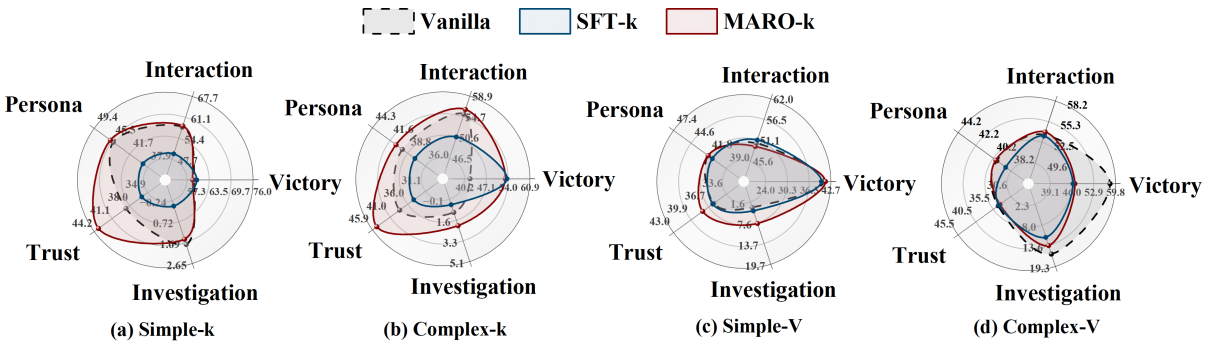

Figure 9: Performance comparison when killer faction is enhanced (**SFT-k**, **MARO-k**). The killer role is controlled by enhanced models while the Victim role is controlled by **Vanilla**.

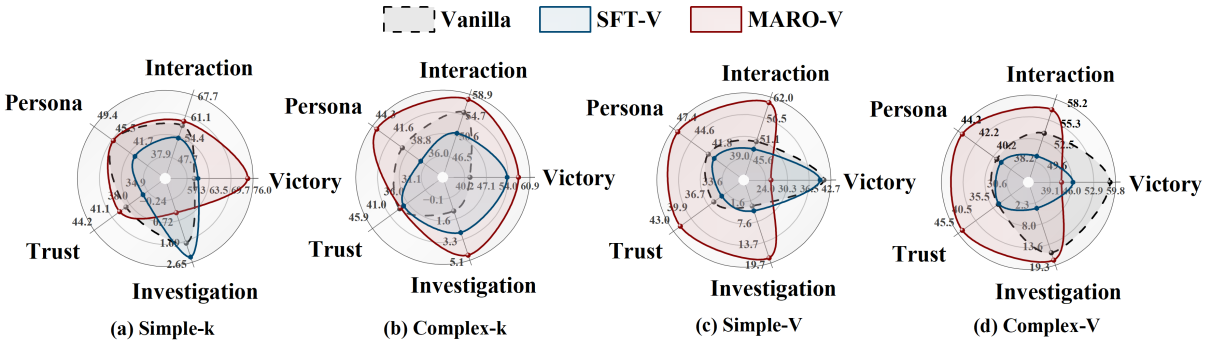

Figure 10: Performance comparison when Victim faction is enhanced (**SFT-V**, **MARO-V**). The Victim role is controlled by enhanced models while the killer role uses **Vanilla**.

smaller coverage areas than Vanilla (dashed line), particularly exhibiting significant contraction in the Interaction, Persona, and Trust dimensions. This performance limitation in our experimental setting may be attributed to SFT's reliance solely on positive samples for training, lacking the contrastive learning mechanisms that could help models distinguish between appropriate and inappropriate social behaviors in complex multi-agent environments.

It should be noted that this finding is specific to our multi-agent social reasoning evaluation framework and does not preclude SFT's effectiveness in other social learning contexts or with different training configurations.

C Additional General Capabilities Results

Finding A2

MARO demonstrates consistent improvements across different mathematical reasoning benchmarks and evaluation methods.

Figure 5 and Figure 6 present comprehensive performance comparisons on mathematical reasoning tasks. On Math-500, MARO-Complex achieves 42.6% accuracy under CoT evaluation, representing a 5.2 percentage point improvement over the baseline's 37.4%. On GSM8K, MARO-Complex improves from 55.4% to 59.4% under CoT method, achieving a 4.0 percentage point enhancement.

Finding A3

Complex social interaction environments significantly enhance instruction following capabilities across all evaluation settings.

Figure 8 shows that models trained in complex environments achieve 72.0% accuracy in the Loose-CoT setting (2.6 percentage point improvement) and 67.8% in Strict-CoT setting (1.2 percentage point improvement). The improvement stems from social interaction environments requiring models to strictly adhere to role settings and game rules, naturally enhancing complex instruction understanding and execution capabilities.

Finding A4

Learning through social simulation achieves optimal performance on moderate-difficulty reasoning tasks.

Figure 7 demonstrates that social simulation learning achieves optimal performance on moderate-difficulty reasoning tasks. Interestingly, both configurations slightly underperform the baseline on the simplest tasks while showing mixed results on the most challenging problems, suggesting that the sophisticated reasoning strategies developed through social interaction are most effective for intermediate-level complexity where strategic thinking and adversarial reasoning skills can be optimally utilized.