Sharing Minds during MARL Training for Enhanced Cooperative LLM Agents

Jiaxuan Gao[1,2] **Yule Wen**[1] **Chao Yu**[1,†] **Yi Wu**[1,2,†]
[1] Tsinghua University [2] Shanghai Qi Zhi Institute
[†] Corresponding Authors
{samjia2000, jxwuyi}@gmail.com

Abstract

LLM agents have shown promising capabilities by adopting advanced reasoning techniques such as Chain-of-Thought (CoT). Incorporating Theory of Mind (ToM) inference, which infers the goals and intentions of teammates, into the reasoning process is proven to be beneficial for enhancing the coordination ability of cooperative LLM agents. This work investigates the impact of explicitly augmenting Theory of Mind (ToM) capabilities during MARL training of LLM agents in multi-agent environments. To enhance ToM capabilities, we introduce a novel technique, Mind-Sharing, which obtains the ground-truth answers for the ToM inference of an agent during centralized training by rewriting the hidden minds of the other agent. Our experiments, conducted in the 2-player version of the cooperative game Hanabi, use the MAPPO as the MARL algorithm and LLaMA-2-7B as the base model. We find that the Mind-Sharing mechanism significantly improves both task performance and sample efficiency in MARL training. Our results reveal enhanced ToM capability, surpassing the ToM inference accuracy of a wide range of models in the self-play setting. Surprisingly, the ToM inference skill learned from self-play also generalizes to the cross-play setting.

1 Introduction

Recently, cooperative agents driven by large language models (LLMs) have shown surprising results by broadly adopting advanced reasoning techniques such as Chain-of-Thought (CoT, Wei et al. (2022)) to exhibit human-like thinking and decision-making capabilities (Qian et al., 2024; Mandi et al., 2024; Park et al., 2023; Agashe et al., 2023; Liu et al., 2023; Xu et al., 2023). Theory of Mind (ToM), the ability to infer the inner mental states of teammates, has been shown to be beneficial for enhancing the coordination ability of LLM agents in multi-agent systems (Kosinski, 2023; Agashe et al., 2023) and human-AI coordination (Ding et al., 2024; Liu et al., 2023). ToM inference can be integrated into the reasoning process by explicitly inferring the intentions or the demands of teammates before performing further reasoning and decision-making.

On the other hand, there is a recent trend to train vision language models (VLM) or large language models (LLMs) with small or medium sizes, e.g., 7B, on decision-making tasks with reinforcement learning (RL) (Xiong et al., 2024; Abdulhai et al., 2023; Hong et al., 2023; Zhou et al., 2024; Zhai et al., 2024), achieving even superior performance than powerful larger LLMs such as GPT-4. Zhai et al. (2024) first fine-tune a 7B base VLM on demonstration data provided by larger LLMs with supervised fine-tuning (SFT) and then adopts RL training to enhance reasoning further, achieving even better performance than commercial LLMs, including GPT-4V and Gemini. Although RL can improve the performance of LLM agents, in practice, we find that the accuracy of ToM inference does not increase after RL training.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In this work, we focus on the question: Would explicitly augmenting the ToM capability during training time enhance the coordination ability of LLM? We carry out our study in a 2-player cooperation game, Hanabi, which features high randomness and partial observability, demanding correctly inferring the intention of the other agent. We use MAPPO as the multi-agent RL algorithm (Yu et al., 2022) and LLaMA-2-7B as the base model. To augment the ToM inference accuracy, we propose a novel technique, *Mind-Sharing*, which obtains the ground-truth answers for the ToM inference of an agent during centralized training by rewriting the hidden minds of the other agent. A supervised fine-tuning loss on the ground-truth ToM answers is added to the MAPPO training loss to guide the LLM learning better ToM inference.

In our evaluation of the Mind-Sharing mechanism, our results show that it significantly boosts task performance and sample efficiency of MARL training, as well as the accuracy of ToM inference in both self-play and cross-play scenarios. By incorporating the Mind-Sharing mechanism into MARL training, fine-tuned LLaMA-2-7B exhibits more self-consistent mental state inference than GPT-4.

2 Methodology

2.1 Experiment Setup

2-Player Hanabi Game. In the 2-player Hanabi game, two players collaborate to create a colorful fireworks display using cards with different colors and numbers. Each player can only see the cards in their partner's hand. Players can either give a hint to their teammates or play/discard a card. To correctly play cards, each player has to accurately interpret the moves of the other agent. Hanabi is challenging due to two key aspects, *high randomness*, which comes from drawing cards and the uncertainty of held cards, and *partial observability*, which occurs since players can not observe their own cards.

LLM-based Multi-Agent Coordination. In our framework, after the LLM agent receives a local observation, including the observed cards and history moves of both players, the LLM agent performs ToM inference to infer the intention behind the move of the partner, reason about the current situation, and finally choose an actionable action. Appendix. A provides more details about the framework.

2.2 Mind-Sharing during Centralized Training

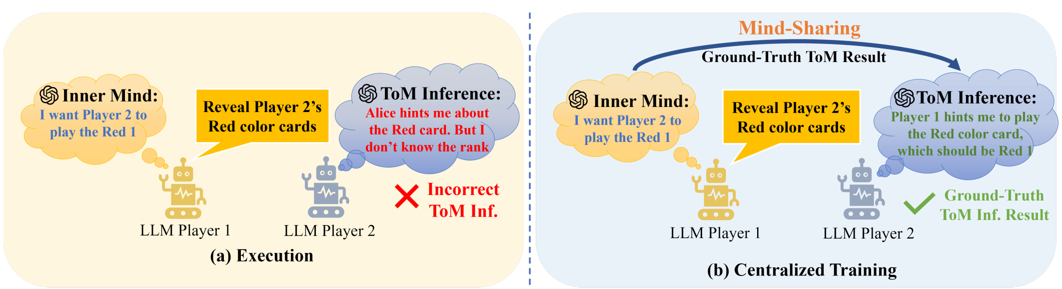

Figure 1: Illustration of the *Mind-Sharing* mechanism. (a) During execution, the LLM Player 1 generates an inner mind before making the decision. The LLM Player 2 then performs ToM inference after observing Alice's move, but this could possibly lead to incorrect ToM inference. (b) During centralized training, we can obtain the ground-truth answer to the ToM inference query of LLM Player 2 by rewriting the inner mind of LLM Player 1.

Fine-tuning LLM Agents with MARL. To fine-tune LLMs on multi-agent coordination tasks, we first perform supervised fine-tuning (SFT) over demonstration reasoning data collected from capable large LLMs and then perform MARL training to enhance the reasoning ability. We use LLaMA-2-7B as the base model and MAPPO (Yu et al., 2022) as the MARL algorithm. The demonstration data are collected with Qwen2-72B-Instruct and processed to avoid catastrophic endings of the trajectories. We follow MAPPO to use a large batch size for MARL training. More details about experiment setup can be found in Appendix. A.

While MARL could enhance the coordination and task reward for cooperative LLM agents in multi-agent environments, updating the LLM solely from reward signals can not reliably improve the ToM ability of the LLM. A key observation is that, though the inner minds produced by the agents are not visible to each other during execution, the inner minds are globally available during centralized training and could be applied to improve the ToM inference accuracy.

Inspired by this observation, we propose the *Mind-Sharing* mechanism, in which we obtain the ground-truth ToM inference answer for an agent by rewriting the inner mind of the other agent. For example, in Fig. 1, LLM Player 1 and LLM Player 2 are playing a game, and the generated inner minds are invisible to each other during the execution phase. In the phase of centralized training, the ground-truth answer to the ToM inference query of LLM Player 2 can be annotated by rewriting the inner mind of LLM Player 1.

To guide LLM with ground-truth ToM inference results, we gather all ToM inference queries and ground-truth ToM inference answers as a dataset $\mathcal{D}_{\text{Mind}}$ and use a supervised fine-tuning loss $\mathcal{L}_{\text{Mind}}$ to guide the LLM to perform ToM inference more accurately,

$$\mathcal{L}_{\text{Mind}} = -\mathbb{E}_{x,y\sim\mathcal{D}_{\text{Mind}}} \log \pi_\theta(y|x) \tag{1}$$

The SFT loss is combined with MAPPO loss during training.

3 Experiments

Evaluation Settings. We perform our experiments under two different settings, Self-Play and Cross-Play. In the self-play setting, two agents supported by identical LLM will cooperate with each other. In the cross-play setting, the two agents are supported by different LLMs. Cross-play evaluation measures the generalization ability of the LLM in coordination tasks.

Evaluation Metric. We consider two metrics, average score and ToM inference error rate. The average score measures the overall performance of an LLM agent in the cooperative Hanabi game. The ToM inference error rate evaluates the ratio of the LLM agent making incorrect ToM inference to the other agent. Specifically, we use an external LLM [1] to judge whether the ToM inference results differ from the ground-truth minds of the other agent. The ToM inference error rate is dedicated to quantifying the ToM capability of different LLMs.

Baselines. We consider various types of baselines, including LLM-based and non-LLM-based ones. For LLM-based baselines, we consider (1) Supervised Fine-Tuning (SFT), which clones the demonstration data with supervised fine-tuning, and (2) Cognitive Architecture for Coordination(CAC) framework proposed by Agashe et al. (2023), which can be supported with different LLMs, specifically GPT-3.5, GPT-4 and Qwen2-72B-Instruct in our experiments. All LLM fine-tuning methods use at most 1M environment samples. For non-LLM-based methods, we follow the training procedure of MAPPO in Hanabi to train an MLP policy that takes vectorized observation as input and outputs atomic actions. This baseline is denoted as MAPPO (non-LLM). We report the scores of MAPPO (non-LLM) with 10M and 100M training samples.

3.1 Self-Play Evaluation

Mind-Sharing can enhance the coordination ability of the LLM-based MA system. Fig.2 shows that MAPPO w. Mind-Sharing gains substantially higher scores than SFT, showing the effectiveness of mind-exchanging centralized training. We also notice that the fine-tuned LLMs outperform the CAC framework supported by various larger language models except GPT-4. Finally, there still exists a large gap between $\text{CAC}_{\text{GPT-4}}$ and MAPPO w. Mind-Sharing, which we believe could be further mitigated with better demonstration data and more environment samples.

Mind-Sharing can induce self-consistent ToM inference. Considering the ToM inference error rate, Fig.2 shows that MAPPO can only achieve nearly the same ToM inference error rate as SFT. This shows that MARL training can not reliably improve the ToM inference skill of the LLM. MAPPO w. Mind-Sharing has the lowest error rate and shows the most self-consistent ToM capability, i.e., the LLM agent can better interpret the behaviors of agents supported by the same LLM. It is worth noting that MAPPO w. Mind-Sharing also makes fewer ToM inference errors than GPT-4, GPT-3.5, and Qwen2-72B-Instruct.

[1] Specifically, we use gpt-3.5-turbo-1106 to evaluate ToM inference error rate.

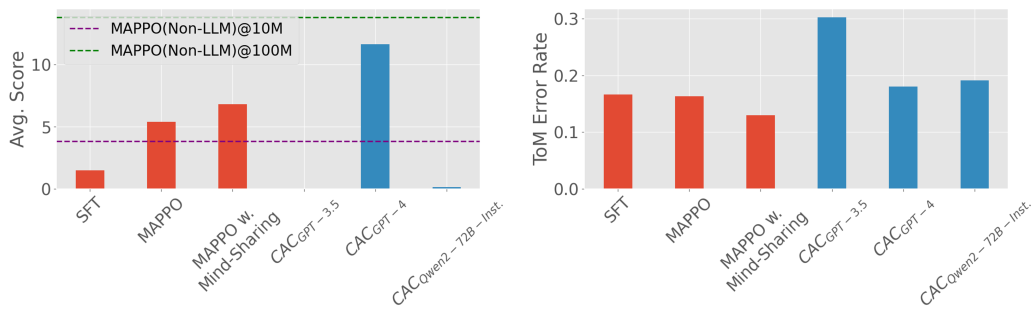

Figure 2: Evaluation in Self-Play Setting. We compare LLaMA-2-7B fine-tuned by various fine-tuning approaches with the CAC agent framework supported by powerful LLMs. Left: Average scores of various approaches. Right: ToM inference error rates in self-play setting. Fine-tuning a 7B LLM with MAPPO w. Mind-Sharing achieves a higher average score than fine-tuning baselines, SFT and MAPPO, and agents supported by advanced LLMs including GPT-3.5 and Qwen2-72B-Instruct, as well as the lowest ToM error rate.

3.2 Cross-Play Evaluation

Mind-Exchange centralized training can enhance the zero-shot coordination ability. In the cross-play setting, we investigate the coordination performance when teaming up one of SFT, MAPPO, and MAPPO w. Mind-Sharing with an expert LLM-based MA system $\text{CAC}_{\text{GPT-4}}$. As shown in Fig.3, SFT and MAPPO achieve similar cross-play scores, and MAPPO w. Mind-Sharing achieves a substantially higher score. This indicates that the policy gradient-based approach can not improve the zero-shot coordination ability, and, by contrast, the Mind-Sharing mechanism can enhance the zero-shot coordination of the LLM agent by improving the ToM inference skill.

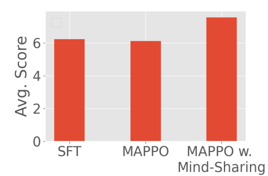

Figure 3: Average Cross-Play Scores with $\text{CAC}_{\text{GPT-4}}$.

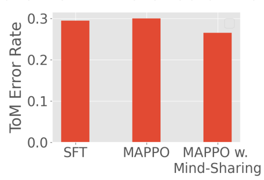

Figure 4: Average ToM error rates of various methods when playing with $\text{CAC}_{\text{GPT-4}}$.

Mind-Exchange centralized training can enhance the ToM capability in cross-play. To further investigate the ToM capability in the cross-play setting, Fig.4 illustrates the ToM inference error rate of SFT, MAPPO, and MAPPO w. Mind-Sharing during cross-play with $\text{CAC}_{\text{GPT-4}}$. MAPPO w. Mind-Sharing is significantly better than SFT and MAPPO in inferring the inner minds of $\text{CAC}_{\text{GPT-4}}$, showing that the learned ToM inference skill generalizes to cross-play setting. This initial experiment shows the potential of an interesting direction to enhance the generalization ability of cooperative LLM agents through self-play.

3.3 Ablation Study

Sample Efficiency. An additional benefit is also illustrated in Fig. 5, where including Mind-Exchange centralized training enhances the sample efficiency of MAPPO .

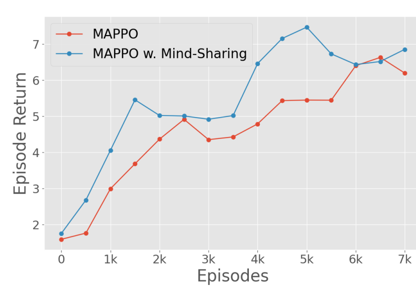

Figure 5: Episode returns of MAPPO and MAPPO w. Mind-Sharing during training.

4 Conclusion

In this work, we investigate the impact of enhancing Theory of Mind inference on MARL training of cooperative LLM agents. To enhance ToM inference, we propose a simple but effective technique, Mind-Sharing. Augmenting ToM inference not only boosts the performance and sample efficiency during training, but also achieves self-consistent ToM inference capability. Furthermore, the learned ToM inference skill also generalizes to cross-play setting, indicating a general improvement of the ToM capability of the LLM. It is an interesting direction to further investigate more environments that involve more agents and language communications.

References

Abdulhai, M., White, I., Snell, C., Sun, C., Hong, J., Zhai, Y., Xu, K., and Levine, S. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. arXiv preprint arXiv:2311.18232, 2023.

Agashe, S., Fan, Y., and Wang, X. E. Evaluating multi-agent coordination abilities in large language models. arXiv preprint arXiv:2310.03903, 2023.

Ding, W., Li, F., Ji, Z., Xue, Z., and Liu, J. Atom-bot: Embodied fulfillment of unspoken human needs with affective theory of mind, 2024. URL https://arxiv.org/abs/2406.08455.

Hong, J., Levine, S., and Dragan, A. Zero-shot goal-directed dialogue via rl on imagined conversations. arXiv preprint arXiv:2311.05584, 2023.

Kosinski, M. Theory of mind might have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083, 2023.

Liu, J., Yu, C., Gao, J., Xie, Y., Liao, Q., Wu, Y., and Wang, Y. Llm-powered hierarchical language agent for real-time human-ai coordination. arXiv preprint arXiv:2312.15224, 2023.

Mandi, Z., Jain, S., and Song, S. Roco: Dialectic multi-robot collaboration with large language models. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 286–299. IEEE, 2024.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology, pp. 1–22, 2023.

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., et al. Chatdev: Communicative agents for software development. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15174–15186, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), Advances in Neural Information Processing Systems, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Xiong, W., Shi, C., Shen, J., Rosenberg, A., Qin, Z., Calandriello, D., Khalman, M., Joshi, R., Piot, B., Saleh, M., et al. Building math agents with multi-turn iterative preference learning. arXiv preprint arXiv:2409.02392, 2024.

Xu, Y., Wang, S., Li, P., Luo, F., Wang, X., Liu, W., and Liu, Y. Exploring large language models for communication games: An empirical study on werewolf. arXiv preprint arXiv:2309.04658, 2023.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. Advances in Neural Information Processing Systems, 35:24611–24624, 2022.

Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. arXiv preprint arXiv:2405.10292, 2024.

Zhou, Y., Zanette, A., Pan, J., Levine, S., and Kumar, A. Archer: Training language model agents via hierarchical multi-turn rl. arXiv preprint arXiv:2402.19446, 2024.

Appendix

A MAPPO Training

In LLM-based multi-agent systems, a typical design choice for the LLM agents is to first infer the mind of the teammates, leverage chain-of-thought reasoning process and finally make the decision. Fig. 6 presents an overview of our LLM-based multi-agent framework. As depicted in Fig. 6, the output of the LLM agent contains three parts, the ToM inference result, the inner mind, and the action, i.e. $v^{\text{out}} = \{v^{\text{ToM}}, v^{\text{mind}}, v^{\text{act}}\}$. A parser is used to extract the actual action from the output text v^{out} to interact with the environment. Our overall training framework is as shown in Fig. 7. To fine-tune the LLM-based MA system, the most straightforward approach is to employ MAPPO to optimize Eq.2 with $\pi_\theta(a_i^t|o_i^t) = \pi_\theta(v_i^{\text{ToM}}, v_i^{\text{mind}}, v_i^{\text{act}}|v_i^{obs})$, i.e. by updating the probability of the output text sequences,

$$\mathcal{L}_{MAPPO} = -\mathbb{E}_{\tau,i}\left[\min\left(\frac{\pi_\theta(v_i^{\text{out},t}|v_i^{\text{obs},t})}{\pi_{old}(v_i^{\text{out},t}|v_i^{\text{obs},t})}\hat{A}_\lambda(a_i^t|s_t), \text{Clip}\left(\frac{\pi_\theta(v_i^{\text{out},t}|v_i^{\text{obs},t})}{\pi_{old}(v_i^{\text{out},t}|v_i^{\text{obs},t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_\lambda(a_i^t|s_t)\right)\right] \tag{2}$$

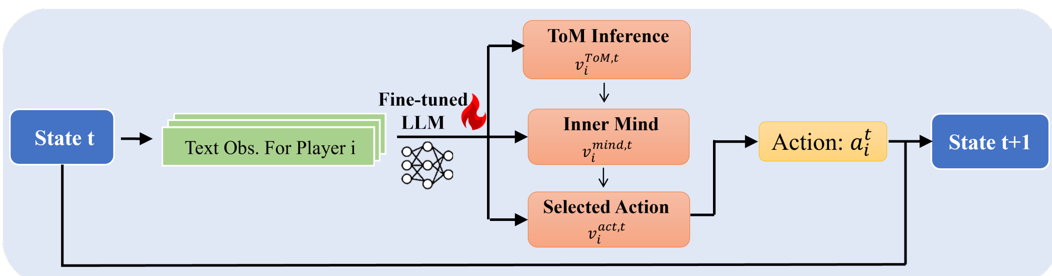

Figure 6: An overview of our LLM-based multi-agent framework. Each agent is supported by a fine-tuned LLM. At each time step t, the acting player i receives text observation from the environment, and performs chain-of-thought reasoning by inferring the mind of the teammates, reasoning the situation as "Inner Mind", and finally selecting the most proper action. A parser is then used to extract the atomic action to interact with the environment.

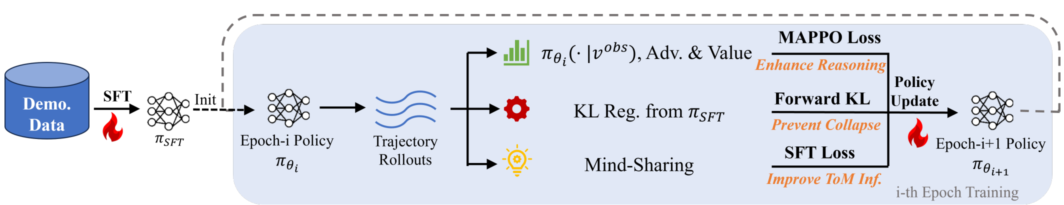

Figure 7: The training procedure. We first run supervised fine-tuning over collected demonstrations to obtain the SFT policy as a starting point. In each epoch of training, we collect trajectory rollouts. Given the trajectories, we compute advantages and values, generate responses with SFT policy, and perform Mind-Sharing to label ground-truth ToM inference answers. Finally, the LLM policy is updated with a MAPPO loss to enhance CoT reasoning, forward KL regularization to avoid model collapse, and an SFT loss to improve ToM inference. Value model is omitted for simplicity.

In the subsequent sections, we will first describe the failure of adopting MAPPO to fine-tune LLM-based MA system and discuss the challenges in Sec.A.1. In Sec.A.2, we discuss techniques that can help resolve the challenges and improve training stability, including choice of KL regularization and advantage shifting.

A.1 Challenges for Training LLM-based Multi-Agent System

Training Setup. Our experiments and studies are taken on the 2-player Hanabi game. To train a strong policy in the 2-player Hanabi game, MAPPO utilizes $1k$ parallel environments with an episode length of 100 to collect online policy rollout data in each epoch, which consumes a significantly larger amount of samples in one epoch than PPO in simpler scenarios does. We follow MAPPO to collect a large number of policy rollouts in each epoch, more specifically, 512 episodes, which contains over 25k samples per epoch. Our experiments use LLaMA-2-7B as the base LLM model and are taken on one node with 8×A100 GPUs. To equip the LLM with basic reasoning and decision-making ability, we run Supervised Fine-Tuning (SFT) over demonstration data collected from Qwen2-72B-Instruct. Since the maximum batch size for fine-tuning the LLM is 128 due to limited computation, we choose to split the large batch of samples into a number of mini-batches, following the practice of MAPPO.

Fig. 5 illustrates the training curve of MAPPO. MAPPO fails to optimize the task reward of the LLM-based multi-agent system, and the performance of the LLM quickly degrades within the first few epochs. To further study the causes of the phenomena, we record the ratio of responses that contain a mixture of random words or characters after the first epoch of update. The probability of the LLM outputing meaningless words significantly increases to 16.5% after one epoch of MAPPO training, indicating a signal of model collapse.

There are several reasons that could lead to the model collapse issue during MAPPO training. First, the model is not properly regularized to the base model or the SFT model. A popular choice in LLM alignment and RLHF is to add a reverse KL term to regularize the LLM towards a reference model, which could help prevent the model from producing meaningless texts. However, as we will show in Sec.A.2, the common choice of reverse KL regularization could not effectively regularize the LLM and forward KL would be more effective for preventing model collapse. Second, while the estimated advantage in MAPPO is believed to reduce the bias for policy update, the negative advantages would cause negative gradients during LLM fine-tuning and increase the probability of nonsense outputs. The issue of negative gradient would be even more severe during the first few epochs since value estimation is highly inaccurate and noisy.

A.2 Towards Stable Training for LLM-based Multi-Agent System

A.2.1 Regularization with KL Divergence

To prevent the LLM from model collapse and generating nonsense outputs during training, the output distribution of the LLM should be regularized to a proper prior distribution, e.g. the distribution of the SFT model π_{SFT}. Here we focus on two common choices, Forward KL regularization $D_{KL}(\pi_\theta||\pi_{SFT})$ and Reverse KL regularization $D_{KL}(\pi_{SFT}||\pi_\theta)$.

Reverse KL regularization is commonly used in RLHF for LLM alignment. To apply reverse KL regularization, we turn to token-level clipping objective and add a token-level reverse KL penalty term to the reward. The objective of MAPPO with reverse KL regularization is,

$$\mathcal{L}_{\text{MAPPO w. Reverse KL}} = -\mathbb{E}_{\tau,t,i,j}[\min(f(v_i^{\text{obs},t}, v_i^{\text{out},t}, j)(\hat{A}_\lambda(a_i^t|s_t) + R_{\text{Reverse KL}}(v_i^{\text{out},t}, j)), \tag{3}$$

$$\text{Clip}\left(f(v_i^{\text{obs},t}, v_i^{\text{out},t}, j), 1-\epsilon, 1+\epsilon\right)(\hat{A}_\lambda(a_i^t|s_t) + R_{\text{Reverse KL}}(v_i^{\text{out},t}, j))] \tag{4}$$

where $f(v_i^{\text{obs},t}, v_i^{\text{out},t}, j)$ is the ratio of probability for token $v_{i,[j]}^{\text{out}_t}$ and $R_{\text{Reverse KL}}(v_i^{\text{out},t}, j)$ is the return of reverse KL penalty,

$$f(v_i^{\text{obs},t}, v_i^{\text{out},t}, j) = \frac{\pi_\theta(v_{i,[j]}^{\text{out},t}|v_i^{\text{obs},t}, v_{i,[0:j-1]}^{\text{out},t})}{\pi_{\theta_{old}}(v_{i,[j]}^{\text{out},t}|v_i^{\text{obs},t}, v_{i,[0:j-1]}^{\text{out},t})} \tag{5}$$

$$R_{\text{Reverse KL}}(v_i^{\text{out},t}, j) = -\sum_{k=j}^{|v_i^{\text{out},t}|-1} \beta \log \frac{\pi_{\theta_{old}}(v_{i,[k]}^{\text{out},t}|v_i^{\text{obs},t}, v_{i,[0:k-1]}^{\text{out},t})}{\pi_{\text{SFT}}(v_{i,[k]}^{\text{out},t}|v_i^{\text{obs},t}, v_{i,[0:k-1]}^{\text{out},t})} \tag{6}$$

Note that the estimated advantage $\hat{A}_\lambda(a_i^t|s_t)$ is state-level and the KL penalty term $R_{\text{Reverse KL}}(v_i^{\text{out},t}, j)$ is token-level, which offers a nice combination of trajectory-level feedback and sentence-level regularization.

In contrast, **Forward KL regularization** is does not modify the reward, but instead combining the MAPPO loss with a reverse KL loss. The objective with forward KL regularization is,

$$\mathcal{L}_{\text{MAPPO w. Forward KL}} = \mathcal{L}_{\text{MAPPO}} + D_{KL}(\pi_{SFT}||\pi_\theta) \tag{7}$$

$$= \mathcal{L}_{\text{MAPPO}} + \mathbb{E}_{\tau,t,i}\left[\mathbb{E}_{y\sim\pi_{SFT}(\cdot|v_i^{\text{obs},t})}\left[\log \pi_{SFT}(y|v_i^{\text{obs},t}) - \log \pi_\theta(y|v_i^{\text{obs},t})\right]\right] \tag{8}$$

Table. 8 reports the forward KL divergence and reverse KL divergence between the SFT model and the LLM model after one epoch of training when using forward KL, reverse KL and no explicit KL regularization. The results indicate that reverse KL can not properly regularize the LLM during training, achieving a close regularization effect as no explicit regularization. On the other hand, forward KL regularization can effectively maintain the both the forward KL and reverse KL divergence between the LLM and the SFT model. This is because the reverse KL penalty in Eq. 4 estimates $D_{KL}(\pi_{\theta_{old}}||\pi_{SFT})$ instead of $D_{KL}(\pi_\theta||\pi_{SFT})$ and during training the policy model π_θ would deviates from the last-epoch policy $\pi_{\theta_{old}}$. To ensure a stable training process, we use forward KL regularization.

	no KL	Reverse KL	Forward KL
Forward KL $D_{KL}(\pi_{SFT}\|\|\pi_{\theta_1})$	3.55	3.51	**2.89**
Reverse KL $D_{KL}(\pi_{\theta_1}\|\|\pi_{SFT})$	4.40	4.21	**3.82**

Figure 8: The forward KL and reverse KL divergence between the SFT model and the 1-st epoch policy π_{θ_1}. Forward KL can regularize the LLM policy the most effectively.

A.2.2 Advantage Shifting

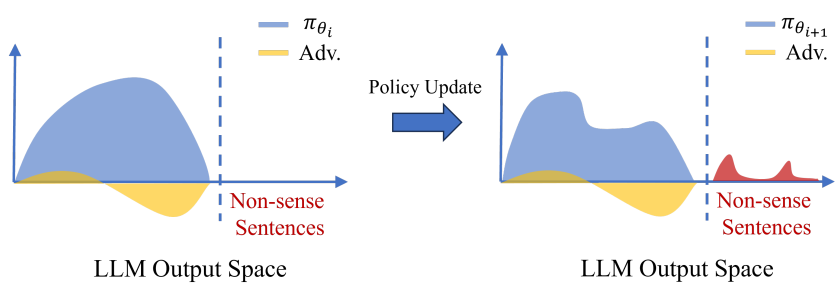

Figure 9: The effect of negative advantages in MAPPO training. While policy update in MAPPO training could decrease the output probability of responses that have negative advantages, the probability of non-sense outputs may increase.

Another issue of vanilla MAPPO is that the negative gradients brought by negative advantages could lead the model to generate nonsense outputs. Fig. 9 illustrates the effect of negative gradients for training LLMs. If the output probability of a generated response y is decreased during training, the output probability of other responses, including the nonsense ones, would possibly increase. The issue is much less severe for RL agent in games because the action spaces of games are usually much smaller than the action space of LLM.

To mitigate the effect of negative gradients, we propose to shift the advantages with a epoch-dependent threshold. More specifically, for the advantages of the policy rollout data in one epoch, $\{Adv_0, Adv_1, \ldots, Adv_{M-1}\}$, we first normalize the advantages,

$$Adv_i^{norm} = (Adv_i - \mu)/\sigma \tag{9}$$

where μ and σ are the mean and standard deviation of $\{Adv_0, Adv_1, \ldots, Adv_{M-1}\}$. Then we shift the normalized advantages by subtracting a threshold T that is the value of the $\lceil \alpha M \rceil$-th smallest advantage value,

$$Adv_i^{shift} = Adv_i^{norm} - T \tag{10}$$

Here we use threshold ratio $\alpha (0 \leq \alpha \leq 1)$ to control the ratio of negative gradients.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our proposed methods and experiment results are shown in Sec. 2.2 amd Sec. 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitations are discussed in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: There are no theoretical results in this paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details are released in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: In the current moment, the code and data are not released yet. We will release the code and data in the near future.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment details aqre provided in the Experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Since evaluation with LLM is costly, we only use perform limited seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: We use 1 node with 8xA100 GPUs to perform our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: skipped

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We believe our method has no significant societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: skipped.**[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:

    Justification: skipped.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: skipped

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Skipped

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Skipped

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.