Machine Intelligence Research
www.mi-research.net
DOI: 10.1007/s11633-025-1547-3

Theory of Mind Inspired Large Reasoning Language Model Improved Multi-agent Reinforcement Learning Algorithm for Robust and Adaptive Partner Modelling

Xiyun Li[1,2] Tielin Zhang[1,3,4] Chenghao Liu[1] Shuang Xu[1,3] Bo Xu[1,2,3]

[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2]School of Future Technology, University of Chinese Academy of Sciences, Beijing 101408, China

[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

[4]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China

Abstract: The cooperative multi-agent reinforcement learning (MARL) field has experienced remarkable progress. However, these advanced methods still face substantial challenges in real-world applications. A significant direction for improving cooperative MARL techniques and addressing existing challenges is robust and adaptive partner modelling. Reasoning about the beliefs of partners, such as their intentions and behaviors, is crucial for partner modelling, which is known as the theory of mind (ToM) in cognitive science. In animals, biological ToM reasoning in the prefrontal cortex (PFC) plays an important role in complex environment survival before decision-making. However, the biological PFC is too complex to be directly incorporated into conventional artificial neural networks (ANNs) in either functional or structural manners. Large reasoning language models (LRMs) have recently demonstrated significant human-like reasoning abilities and impressive performance. Therefore, we propose an improved LRM framework to simulate the PFC for robust and adaptive partner modelling. Despite the excellent performance of LRMs in various fields, their ToM reasoning capabilities remain limited in complex MARL scenarios. Therefore, we further propose a ToM reasoner to enhance the ToM reasoning abilities of LRMs. Our framework exhibits robustness and adaptability across various LRM sizes, improving the ToM reasoning ability of agents and facilitating more effective partner modelling, thereby achieving higher performance scores in cooperative benchmarks.

Keywords: Theory of mind, multi-agent reinforcement learning, partner modelling, large reasoning language model, biological decision-making model.

Citation: X. Li, T. Zhang, C. Liu, S. Xu, B. Xu. Theory of mind inspired large reasoning language model improved multi-agent reinforcement learning algorithm for robust and adaptive partner modelling. *Machine Intelligence Research*. http://doi.org/10.1007/s11633-025-1547-3

1 Introduction

Cooperative multi-agent reinforcement learning (MARL) has attracted considerable attention because of its potential to coordinate multiple agents to achieve common goals in complex environments[1–3]. However, cooperative MARL scenarios face various challenges, such as nonstationary and credit assignment challenges, requiring agents to break symmetry and cooperate efficiently. Several MARL approaches[4–7] have been developed to address these challenges. However, these methods overlook partner modelling before decision-making, potentially hindering their practical applications in terms of performance and scalability. Therefore, establishing robust and adaptive partner modelling for the MARL algorithm to accurately predict and estimate the behaviors and intentions of other agents is crucial for efficient cooperation.

Modelling partners in cooperative MARL tasks is a special case of opponent modelling[8–11], a crucial research direction in multi-agent systems for solving the non-stationarity challenge. Traditional opponent modelling methods include strategy reconstruction, type reasoning, intention recognition, recursive reasoning and other methods[8]. However, these approaches have drawbacks, including adaptation gaps, complex feature engineering and insufficient state-space representation capabilities. Researchers have attempted to address these critical challenges through explicit partner modelling. Some methods con-

Research Article
Manuscript received on May 31, 2024; accepted on January 17, 2025
Recommended by Associate Editor Dacheng Tao
Colored figures are available in the online version at https://link.springer.com/journal/11633

struct additional behavior models to simulate partner agents involving analyzing opponents′ context information, such as historical trajectories, to characterize their behaviors and predict their actions[12, 13]. Although explicit partner modelling methods can model partners from some perspectives, they often suffer from limited partner reasoning, interpretability and a lack of interaction capabilities.

Large language models (LLMs), such as GPT-4 and GPT-o1[14], have recently shown expert-level abilities across various areas, profoundly influencing people′s lives[15–18]. This development has spurred interest in leveraging LLMs′ advanced reasoning capabilities to drive innovation and accelerate progress in various fields, from code generation[19, 20] to human-like autonomous agents[21–24]. Recent studies have demonstrated that integrating code datasets and chain-of-thought (CoT) prompting significantly enhances the reasoning capabilities of LLMs[25, 26]. LLMs with strong reasoning capabilities can be denoted as large reasoning language models (LRMs). In this work, we utilize an LRM to construct a robust and adaptive partner modelling framework, enhancing the partner reasoning ability and interpretability of MARL methods.

The biological social decision-making model is one of the most critical theories in neuroscience, as it explains why humans can achieve efficient social cooperation. This model comprises two key components: intuitive decision-making driven by reinforcement learning and reasoning decision-making based on belief-based learning, which involves anticipating the intentions and actions of others[27, 28]. Humans integrate values from intuitive decision-making (e.g., goals) and beliefs from social reasoning (e.g., intentions of others) to achieve efficient social decision-making. Research indicates that belief-based learning is primarily associated with the prefrontal cortex (PFC) region[29], making it crucial to construct a model that simulates the belief-based learning mechanism in the PFC for human-like partner reasoning.

The PFC plays a critical role in belief-based learning by performing mental reasoning, also known as theory of mind (ToM). Therefore, we develop a ToM reasoning module to simulate the PFC for partner modelling. ToM[30–33] is a crucial psychological concept that emphasizes people′s ability to understand and reason about the goals, intentions and mental states of others. The incorporation of ToM into partner modelling in the cooperative MARL is a promising research direction that will facilitate efficient collaboration. Recent computational models of ToM have facilitated value alignment between humans and agents[34] and fostered efficient communication among multiple agents[35].

Inspired by the biological social decision-making model, we propose a biologically plausible LRM-improved MARL (LRM-MARL) framework to further enhance adaptive partner modelling and efficient cooperation. In our framework, we propose a ToM reasoning module to improve the mental reasoning capabilities of the LRM for cooperative MARL tasks. To validate the ToM ability of our framework, we have conducted diverse experiments on the basis of our previous work[36]. The experimental results demonstrate the effectiveness of our framework in mental reasoning and partner modelling, which significantly enhances competitive MARL methods. Our contributions can be summarized as follows:

1) Inspired by the biological social decision-making model, we design a partner modelling LRM-MARL framework. Our framework incorporates a ToM reasoner module as the belief learning component for better ToM reasoning ability, which comprises an information extractor, our LRM, and the LRM augmenting module (LAM).

2) Among a large number of recently developed LRMs, we select CodeGen in our ToM reasoning module to simulate the PFC. To validate the biological plausibility of our ToM reasoning module as the PFC, we conduct analyses from structural and functional perspectives and construct diverse experiments for further verification.

3) The experimental results demonstrate that our proposed partner modelling framework exhibits superior reasoning capabilities and cooperative performance across various maps, indicating that our framework can successfully explain the beliefs of partners and improve collaboration efficiency in MARL cooperative tasks.

2 Related works

2.1 Cooperative multi-agent reinforcement learning

Significant breakthroughs have been recently made in cooperative MARL, facilitated by the development of many advanced networks and MARL techniques[4, 5, 7]. These methods are commonly categorized into two categories: value-based methods and policy-based methods. Independent Q-learning (IQL)[37] extends the deep Q-network (DQN) paradigm to cooperative MARL, which involves the interaction between two learning agents. Q-value mixing (QMIX)[4], a value-based approach, integrates the centralized training decentralized execution framework and a mixing network to estimate joint action values as a monotonic combination of individual agent values. The actor-critic method counterfactual multi-agent policy gradient (COMA)[5] approaches the credit assignment challenge by leveraging counterfactual baselines. Qtran[6] attempts to enhance QMIX by alleviating certain structural constraints. However, these methods overlook the construction of ToM models for other agents, which are crucial for inferring their intentions and predicting their subsequent actions.

2.2 Partner modelling

Understanding and predicting the actions and inten-

tions of partner agents is crucial for achieving efficient cooperation in multi-agent scenarios. Some existing research on partner modelling focuses on characterizing the styles and strategies of partners and predicting their actions in MARL scenarios. Recent work[9] utilizes the maximum entropy method for partner modelling. He et al.[10] construct an additional partner modelling module to estimate the q-tables of partner agents. Wen et al.[11] propose a partner modelling method through multi-step recursive reasoning. Inspired by cognitive science, Li et al.[12] predict partner styles by constructing a ToM module, achieving efficient cooperation with unseen partners in the cooperative overcooked environment. Wu et al.[38] focus on implicit modelling in interactions with various opponents or partners. Nonetheless, in contrast to the ToM cognitive process in humans, which incorporates substantial common-sense world knowledge and task-specific prior knowledge, these partner modelling modules do not consider such knowledge, limiting their effectiveness and interpretability.

2.3 Large reasoning language model

The exponential growth of LRMs[14, 17] has profoundly impacted various industries to address challenging tasks[15, 16, 18, 21, 22]. Some LRMs demonstrate exceptional reasoning skills in understanding intricate linguistic structures and making accurate decisions[25, 26]. These reasoning abilities of LRMs can be elicited through the CoT prompting equipment[26, 39] and the extensive code corpora, which guide the model in thinking step by step[19, 40]. While existing LRMs demonstrate exceptional overall abilities, they still lack sufficient ToM reasoning capabilities for efficient collaboration[41], requiring the integration of various cognitive skills. Our research aims to design a framework to enhance the ToM reasoning capabilities of existing LRMs for partner modelling in multi-agent tasks, achieving efficient cooperation among agents.

2.4 Theory of mind in MARL

ToM[30, 42, 43] is a crucial cognitive ability that allows individuals to perceive, comprehend and attribute unobservable mental states of others, such as thoughts, desires and emotions[31–33, 43]. The ToM ability facilitates social interactions, communication, empathy, self-awareness and moral reasoning, fostering human accomplishments. Therefore, researchers have endeavoured to equip AI agents with ToM capabilities to address critical challenges in MARL, such as low sample efficiency and poor generalizability[34, 35, 44, 45]. The ToM module predicts the values and intents of human users based on their instructions and feedback for effective bidirectional human-robot communications[34]. Wang et al.[35] utilize the ToM module to anticipate the priority of communication between agents, achieving more efficient agent communication and cooperation. By leveraging the historical trajectory data of other agents, agents with ToM capabilities can forecast their subsequent trajectories[44]. Some researchers constructed mental models for the human-robot teaming[45]. In contrast to previous works, our work focuses primarily on the reasoning aspect of ToM. Through our ToM reasoning module, agents can reason and infer the intentions, goals and actions of other agents, thereby facilitating efficient collaboration.

3 Methodology

3.1 Problem definition

In cooperative multi-agent problems, a 2-player Markov decision process can be defined as a tuple $\left(\mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{M}, \mathcal{Q}, \gamma, \rho^i\right)$, where $\mathcal{O}$ represents the observation space and $\mathcal{A}$ represents the action space shared by both the ego and partner agents. The joint action for the ego and partner agents can be denoted by $\mathbf{a} = (a^1, a^2)$, whereas the joint observation can be represented as $\mathbf{o} = (o^1, o^2)$ consisting of the ego and partner observations. $\mathcal{P}$ defines the environment transition probability function $\mathcal{P} : \mathcal{O} \times \mathcal{A} \to \mathcal{O}$. In our experiment, the ego and partner agents share the same structure and reward function. The reward function $\mathcal{R} : \mathcal{O} \times \mathcal{A} \to \mathbb{R}$ is the same for all the agents.

The ToM reasoner model $\mathcal{M}$ can reason for the partner agent on the basis of the historical context information of the partner agent and observation information at the current time step t. $\mathcal{Q}$ denotes the partner reasoning space and $\gamma \in [0, 1)$ is the discount factor used for future rewards. At time step t, the ego agent perceives environmental observation $o_t^1 \in \mathcal{O}$ and obtains the ToM partner reasoning $q_t^1 \in \mathcal{Q}$ from the ToM reasoner $\mathcal{M}$, taking action $a_t^1 \in \mathcal{A}$ drawn from the ego policy $\rho^1 : \mathcal{O} \times \mathcal{A} \to [0, 1]$, denoted as $a_t^1 = \rho^1\left(\cdot \mid o_t^1, q_t^1\right)$. The partner policy can be denoted as $a_t^2 = \rho^2\left(\cdot \mid o_t^2, q_t^2\right)$. The ego and partner agents transit to the next state $\mathbf{o}_{t+1}$ with probability $\mathcal{P}\left(\mathbf{o}_{t+1} \mid \mathbf{o}_t, \mathbf{a_t}\right)$, receiving a numerical reward r_{t+1} from the environment. Agents aim to maximize the cumulative discounted return $\sum_t \gamma^t r(\mathbf{o}_t, \mathbf{a}_t)$ via efficient collaboration. The detailed model structure is presented in Section 3.3.

3.2 Biological decision-making structure

The social decision-making model in neuroscience suggests that human decision-making predominantly relies on two mechanisms: intuitive decision-making, which is based on reinforcement learning through trial and error, and ToM reasoning decision-making, which is based on belief learning and involves predicting and anticipating the actions of others[27, 28, 46–48]. To refine these two mechanisms, we summarized a biological decision-making

structure in the brain based on analysis from relevant interdisciplinary papers[49–52], as depicted in Fig. 1. The structure comprises two distinct pathways[46–48]: an intuitive decision-making pathway and a ToM reasoning pathway. The intuitive pathway forms a fast decision-making system, including the sensory region and basal ganglia (BG). The ToM reasoning pathway, which forms the slow decision-making system, involves cooperation among multiple brain areas, including the sensory region, BG, medial prefrontal cortex (mPFC) and dorsolateral prefrontal cortex (dlPFC)[49, 50].

The sensory region in the brain plays a critical role in processing environmental observations and extracting crucial features, which is fundamental for ToM reasoning and decision-making. This region processes environmental observations and transmits essential encoded information to the PFC, an important area in cognitive control, with the ability to orchestrate thought and action according to the goals[52]. The mPFC area in the PFC is crucial for interpreting environmental cues and constructing the reasoning representations of others[51, 52]. The mPFC then translates this sensory input into ToM reasoning, which is conveyed to the dlPFC. The dlPFC refines this input and plays a supramodal role in various executive functions, including attention selection, working memory, intricate partner reasoning and belief-making. It can adapt to environmental changes, collaborating with other regions to increase decision-making efficiency[53]. Therefore, the dlPFC maintains strong connections with the mPFC and the BG, which is crucial for modulating mental representation and generating partner state-action reasoning for decision-making processes[54]. Finally, partner reasoning reaches the BG, which comprises some subcortical nuclei essential for regulating motor and cognitive functions, including attention and decision-making[49, 50, 55]. These brain regions dynamically interact within human cognition to manage input information, facilitate ToM reasoning and make decisions.

3.3 LRM-improved MARL framework

Inspired by the biological decision-making structure depicted in Fig. 1, we have developed our LRM-MARL framework to enhance ToM reasoning and facilitate efficient collaboration on the basis of our previous work[36]. As illustrated in Fig. 2, our proposed framework comprises two core modules: An MARL module for decision-making, which mirrors the intuitive decision-making pathway in the biological decision-making structure, and a partner modelling ToM reasoning module (ToM reasoner) based on the LRM for belief learning, which simulates the ToM reasoning pathway. This framework aims to enhance the ToM ability of agents in the decision-making process by combining the strengths of intuitive decision-making and ToM reasoning.

As shown in Fig. 2, our framework utilizes the observation encoder to preprocess and gather information from environmental observations o_t^i to generate the environment embedding e_t^i. The observation o_t^i of agents comprises the information of each grid node within the visible range of agents, including the positions of agents, as well as observable features such as keys, locks and diverse terrains. As shown in (1), we have developed the ToM reasoner as a partner modelling module, with environment embedding e_t^i as its input. The MARL component in our framework can employ various MARL methods, including both value-based and policy-based approaches, such as QMIX and COMA. As shown in (2), the partner modelling in our ToM reasoner component includes three stages: extracting information, ToM reason-

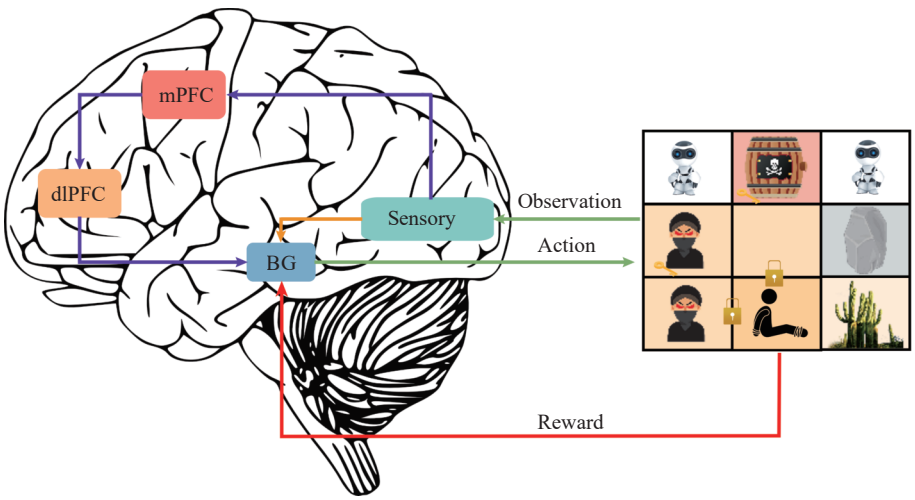

Fig. 1 The biological decision-making structure, including an intuitive decision-making pathway (orange) and a ToM reasoning pathway (purple). The green and red pathways represent the interaction process between the human and the environment, and the environmental reward feedback, respectively. (Colored figure is available in the online version at https://link.springer.com/journal/11633)

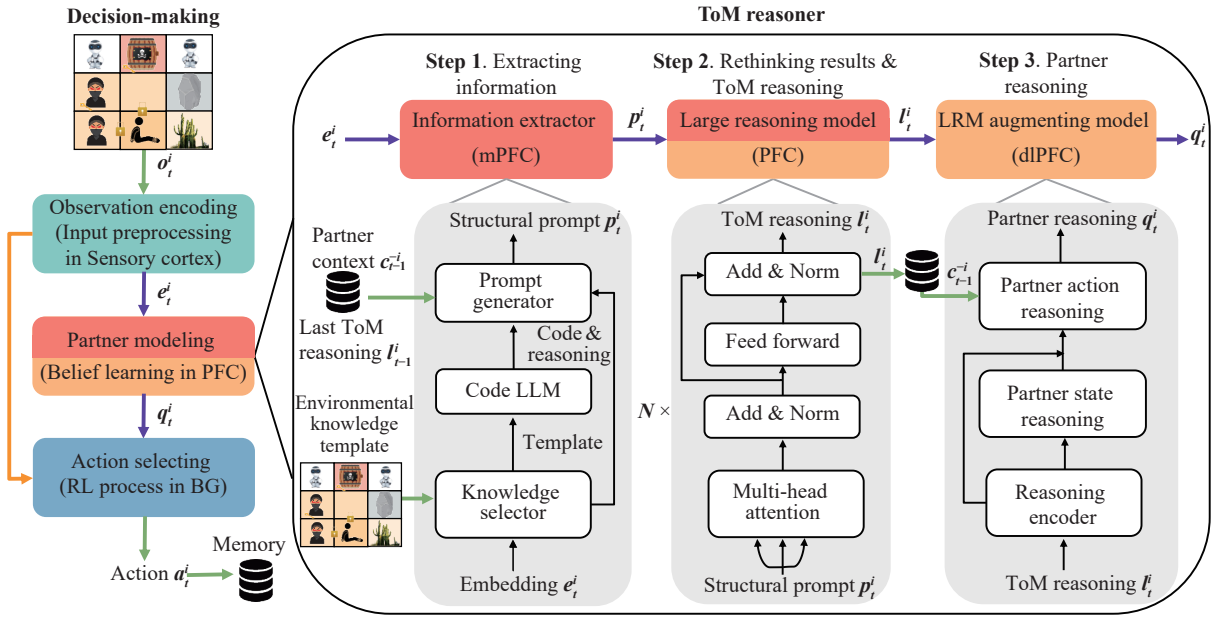

Fig. 2 Our proposed LRM-MARL framework is inspired by the biological decision-making structure, comprising an intuitive decision-making pathway based on the MARL module and a ToM reasoner for partner belief learning. The ToM reasoner for partner modelling consists of three stages: extracting information, rethinking results and ToM reasoning, and partner reasoning. These stages are completed through an information extractor module, the LRM model and the LRM augmenting module.

ing and partner reasoning in the cooperative MARL scenarios. It comprises an information extractor, the LRM, and an LRM augmenting module (LAM).

In our ToM reasoning module, the first stage involves information extraction by an information extractor, which is composed of three parts. First, a template-based comment generation function serves as the knowledge selector, utilizing the partner context c_t^{-i} and environment embedding e_t^i to generate the knowledge template formatted as code comments as output. The second part of the information extractor uses a code LLM to generate code reasoning, enhancing the ToM reasoning capabilities of our framework. In the third part, we integrate these knowledge templates with the code-form reasoning and the latest ToM reasoning from the memory to construct a structured prompt p_t^i in the prompt generator. As shown in Fig. 3, our knowledge template includes the environmental descriptions and the rules of our cooperative task.

$$\begin{aligned}&\text{Observation encoding}: e_t^i = \text{Embed}(o_t^i)\\&\text{Partner modelling}: q_t^i = \text{ToM}_{\text{partner}}(e_t^i, c_{t-1}^{-i}, l_{t-1}^i)\\&\text{Action selection}: a_t^i = \rho^i(o_t^i, q_t^i).\end{aligned} \tag{1}$$

Our proposed ToM reasoner component in our framework uses LRM to understand the last action of the partner and achieve ToM reasoning for the beliefs of the partner in the second stage, which is crucial for improving the ToM reasoning ability and partner modelling of agents and achieving more efficient cooperation. The LRM takes p_t^i as input and generates the comprehensive ToM reasoning representation l_t^i, an embedding vector incorporating partner reasoning. By using this LRM, we can generate an interpretable textual output for the partner and environment.

$$\begin{aligned}&\text{Extracting information}: p_t^i = f_{\text{mPFC}}(e_t^i, c_{t-1}^{-i}, l_{t-1}^i)\\&\text{ToM reasoning}: l_t^i = f_{\text{PFC}}(p_t^i)\\&\text{Partner reasoning}: q_t^i = f_{\text{dlPFC}}(l_t^i, c_{t-1}^{-i}).\end{aligned} \tag{2}$$

To bridge the gap between the semantic space of LRM and the state-action space in MARL, we introduce an additional LAM module in the third stage. The LAM module contains two pathways, mapping the ToM representation l_t^i to the partner state and action space, respectively. Input of the LAM includes l_t^i and the partner context c_t^{-i} from memory, and the output is the partner reasoning representation q_t^i. We concatenate q_t^i with the environmental observation o_t^i as input to the MARL methods for the decision-making process.

3.4 Biologically plausible ToM reasoner

Existing research underscores the pivotal role of code in enhancing the reasoning abilities of LRMs[26, 39]. Therefore, we choose Codegen, a code corpus-based language model[20], to simulate the mPFC in our proposed partner modelling ToM reasoner module. Codegen represents a significant advancement in program synthesis LLMs[20], which is trained on both natural language and programming language data, and open-sourcing the training library JAXFORMER. Our partner modelling ToM reasoner module employs a transformer-based architecture to

Prompt

Assuming you are a **helpful AI assistant with theory of mind ability**. Your advanced capabilities enable you to process and understand the cooperative task rules, environmental state, partner context information, and other relevant information for achieving ToM reasoning and partner modeling about your partner. Now you can assist the policy of ego agent for making the optimal action.
<$**environment_introduction**$>:
Environment information: There are an ego agent, a partner agent, keys, locks, bandits, an explosive, and a hostage goal in our cooperative task. Terrain variations comprise standard areas, rock obstacles, and cactus areas. Each agent can select from **five actions**: up, down, left, right, stay, denoted by the numerical values 0−4, respectively: 0-down, 1-up, 2-left, 3-right, 4-stay.
Environment rule: The primary objective of agents is to: **reach the hostage location and liberate the hostage** in the camp of bandit criminals. Now this is a big version with a 5×5 grid layout. The ego agent and partner agent do not know the location of the explosive, and reaching it simultaneously is unfeasible. Different terrains have different speeds. The faster the ego agent and partner agent reach the target point simultaneously, the higher the reward will be.
<$**current_observation**$>:
Initialize a zero matrix for the bandit camp, denoted as **5×5 grids *g***
The grid cell with the hostage in the matrix will be assigned a value of 1 and the hostage location is in the position [4,2]. The ego agent grid cell in the matrix will be assigned a value of 2, and the partner agent grid cell in the matrix will be assigned a value of 3.
g = np.zeros((5, 5)), *g*[4, 2] = 1
…

Fig. 3 A prompt example, which is a textual description of the environment generated by the prompt extractor at stage one. The prompt generator combines the knowledge template with the code reasoning as the output of the prompt extractor. The highlighted part represents the code reasoning from the code LLM, whereas the other part represents the knowledge template of the knowledge selector.

achieve high-quality understanding and reasoning from natural language prompts.

In Section 4, we demonstrate the biological plausibility of the ToM reasoner to simulate PFC reasoning by analyzing the multiscale similarity between the PFC and our ToM reasoner. We first validate the similarity from both structural and functional perspectives, followed by multiscale experiments for further validation.

From a structural standpoint, the PFC comprises interlaminar mini-columns[56], a configuration corresponding to the multi-layered structure inherent to the transformer architecture within the ToM reasoner. Similarly, recent research[57] has proposed that PFC regions can make probabilistic inferences about the reliability of the current behavioral strategy and several alternative strategies, thus deciding whether to exploit the existing strategy or explore new strategies. Furthermore, it has been shown[58] that the PFC is vital for verbal analogical reasoning, and other research[56] has highlighted its role in the executive control of task-related target selection and decision-making during a visuomotor delayed match to sample (DMS) task. Therefore, the PFC is crucial for social reasoning and decision-making, a function similar to the robust ToM and reasoning capabilities exhibited by our proposed ToM reasoner.

Our comprehensive analysis demonstrates the multiscale similarity and striking equivalence between the PFC and the ToM reasoner, indicating that the biologically plausible ToM reasoner can serve as a PFC in facilitating human-like cooperation and decision-making among agents. Furthermore, the multiscale similarities between the transformer structure in the ToM reasoner and interlaminar mini-columns in the PFC suggest that the computational experiments of our framework provide insights into human cognitive processes. In Section 4, we will further validate the biological plausibility through two experimental scales: cognitive tests and cooperative MARL tasks. In the scale-1 experiment, we construct cognitive tests, including logical and ToM reasoning tests, to evaluate the reasoning ability of our proposed ToM reasoner. In the scale-2 experiment, we introduce a new collaborative MARL environment, Reason, to verify our framework for partner modelling and efficient collaboration.

4 Experiments

4.1 Cognitive test of ToM reasoner

In Section 4, we designed multiscale experiments to evaluate our proposed partner modelling framework. In the scale-1 experiment, we construct cognitive tests comprising logical and ToM reasoning questions to evaluate the reasoning ability of our ToM reasoner. In the scale-2 experiment, we introduce a new collaborative MARL environment called Reason to verify our framework for partner modelling and efficient collaboration.

As shown in Figs. 4 and 5, we have developed a cognitive test for our scale-1 experiment to verify the multiscale similarity between the PFC and our ToM reasoner from the perspectives of both logical and ToM reasoning on the basis of relevant datasets from previous work[59]. The logical reasoning part spans different difficulty levels and contains intelligence test questions related to mathematical reasoning and pattern recognition.

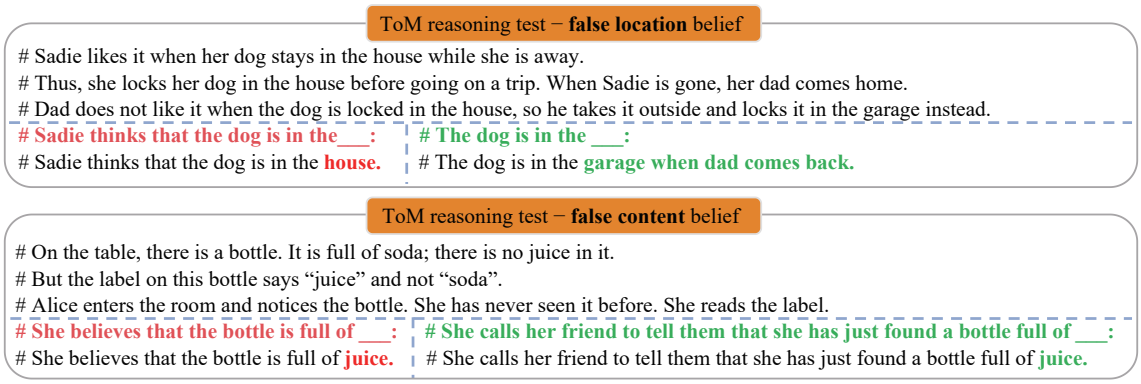

Fig. 4 Examples of the false belief tests: the false location test and the false content test. Each false belief test in our ToM reasoning tests comprises two questions from distinct viewpoints, necessitating accurate responses to all the questions for successful completion. In these examples, the highlighted portions in green and red denote the questions and responses generated by our ToM reasoner, indicating its ToM reasoning ability. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

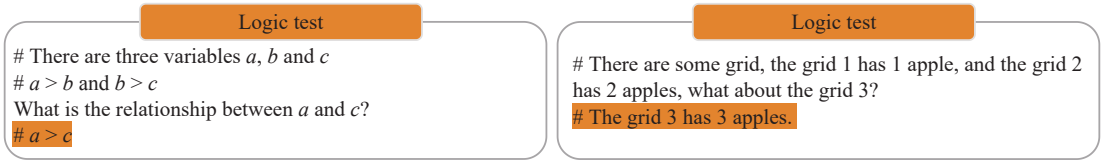

Fig. 5 Examples of logic reasoning tests in our cognitive tests. The **highlighted section** represents the output of the ToM reasoner, whereas the white section represents the question for the ToM reasoner.

The ToM reasoning part in our cognitive tests includes several standard cognitive tests, such as false belief tasks[59, 60], which verify the ToM ability of ToM reasoning to understand other agents.

The false belief test, also known as the Sally-Anne test[61], is a commonly used standard tool in cognitive science for examining the development of children's ToM through their prediction of other people's beliefs. In this test, the researcher presents two dolls to the children, Sally (beside a basket) and Anne (beside a box). Sally put a small ball into the basket, covered it with a cloth, and then Sally left. After Sally left, Anne took the ball out of the basket and put it in the box beside her. After a while Sally came back. At this time, the researcher asked the children, "where will Sally go to find the ball?" This task tests the children's ability in belief reasoning. In cognitive science, belief reasoning is a crucial component of ToM reasoning ability, which is the core ability for partner modelling. Therefore, we can verify the ToM ability of our ToM reasoner via the false belief test.

The ToM reasoner in our proposed partner modelling framework completes our cognitive tests, which demonstrate logical reasoning and strong ToM ability. Some research[62] in the field of cognitive science has indicated that a close link between the mPFC and the dlPFC with the false belief test. Researcher[62] has led to an fMRI experiment to verify and explore this relationship, indicating the important role of the PFC in stimulus-independent mental processes during false belief reasoning, facilitating the shift in attention between stimulus-oriented and stimulus-independent mental processes. Therefore, our experimental results further validate the multiscale similarity between the ToM reasoner and the PFC. In the scale-2 experiment, we apply our proposed partner modelling framework to cooperative MARL tasks, simulating the PFC in the biological decision-making structure to enhance the ToM reasoning abilities of agents.

4.2 Reason environment

As illustrated in Fig. 6, we present our cooperative environment in the scale-2 experiment, Reason, comprising rock obstacles, a target hostage location, some bandits, an explosive zone and two agents. In the Reason environment, the agents need to reason and collaboratively complete a series of subtasks to rescue the hostage from the bandits as quickly as possible. Their tasks involve navigating through an explosive zone, avoiding bandits and dangerous obstacles, and collecting keys to unlock to rescue the hostage. The action space for the two agents includes five distinct actions: moving upwards, downwards, leftwards, rightwards, and remaining stationary. The environmental observation comprises all observable grids, with each grid represented by a high-dimensional embedding vector.

To evaluate the adaptability of our partner modelling LRM-MARL framework across various scenarios, we have constructed various maps with different settings, representing distinct complexity levels. Within the expansive large-scale map of our environment, we incorporate

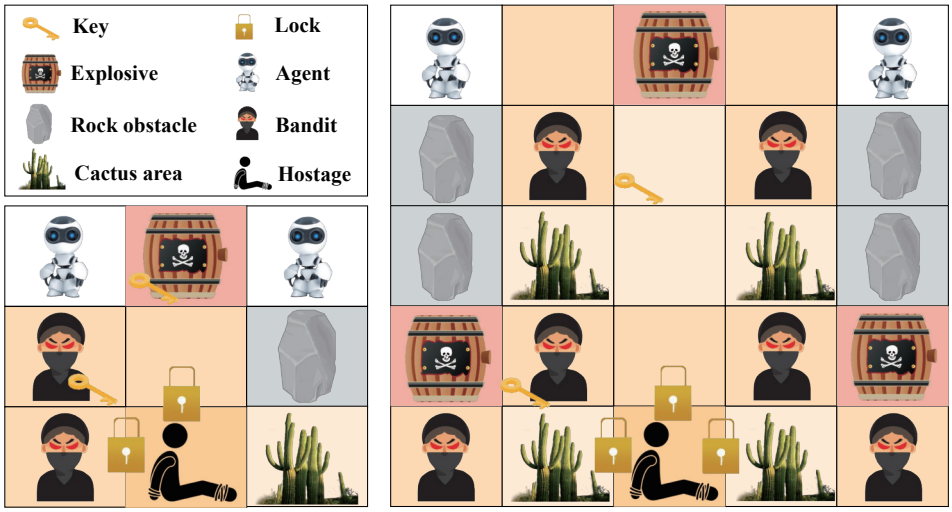

Fig. 6 Overview of the example maps of our reasoning environment (Reason). Our environment consists of two agents, one hostage, explosives, some bandits, the cactus areas, keys and locks and obstacles. The goal of agents is to reach the hostage location and rescue the hostage from the camp of bandits. There are two different map sizes, with the larger maps having more obstacles, a larger exploration space and greater complexity.

broader spatial dimensions, increased obstacle and bandit density, and more terrain types to evaluate our proposed framework′s reasoning capabilities comprehensively.

The explosive zone: To prevent the agents from rescuing hostages, the bandits place weight-sensitive explosives at undisclosed locations unknown to the agents. The explosives are triggered when two agents simultaneously arrive at the explosive′s location. As the observational data of agents do not include specific details about the explosives, such as their exact positions, they must explore and collaborate effectively to navigate through these explosive zones. If the explosives are triggered, agents will return to their starting positions and penalize them. Therefore, strong collaboration is essential for the agents to reach the hostage location successfully.

The obstacles and the bandit zones: In this task, agents are prohibited from entering areas with rock obstacles. Entering the bandit zones significantly increases the time that agents take to complete this rescue task, leading to mission failure. Consequently, entering either obstacle or bandit zones will incur specific penalties for the agents and slow their speed. The environmental information available to the agents includes the precise locations of obstacles and bandits, requiring them to strategically deduce the optimal path by considering both the observational data and the context information of their partners.

The reward criteria: The reward from the environment serves as the principal metric for evaluating the ToM reasoning and cooperation efficiency of the agents. To achieve a high reward, agents must collaborate effectively to accomplish the task as quickly as possible, as longer completion time results in higher penalties. Agents win a high reward when both agents successfully reach the hostage location. A better reward performance requires agents to complete the rescue task within a limited number of steps, demanding advanced reasoning capabilities and effective collaboration.

The ToM reasoning ability of agents is crucial for achieving the mission objective of reaching the target location and rescuing the hostage. Without this ability to comprehend and predict their partners, agents face the risk of becoming trapped in repetitive patterns. As depicted in Fig. 6, if both agents choose to remain stationary and wait for their partner′s move, they will receive penalties at each time step. Alternatively, if the left agent consistently moves right while the right agent moves left, they will not only be reset to their starting positions but also receive additional penalties. Therefore, agents need ToM reasoning ability to break the symmetrical pattern in our task.

The Reason task is complex for several reasons. First, this task is characterized by a highly sparse reward, presenting a significant challenge for MARL methods. Second, unlike conventional search tasks where revisiting a cell is prohibited, agents in this task can revisit all grid cells, mirroring real-world scenarios. Finally, this task environment is partially observable and requires strong collaboration among agents, with unknown locations of explosives and no communication between agents, increasing the task complexity.

4.3 Environmental settings and baselines

For the scale-2 cooperative tasks, our experiment runs for 5 000 epochs for QMIX. Owing the fast convergence speed of COMA, the number of epochs is 200 for the COMA experiments. To accurately assess the effectiveness of our method, we conduct experiments across ten different seeds (0–9), yielding average performance and

variance results. We use the RMSprop optimizer in these methods, and the learning rate is 0.000 5. The reward discount factor is $\gamma = 0.99$, and the maximum length for an episode is 50. The maximum size of the replay buffer is 5 000. We employ gradient clipping to prevent exploding and vanishing gradients. All the experiments were conducted on an AMD EPYC 7 742 server with a single NVIDIA-A100 GPU that can meet our method's computational requirements.

The LRM in our ToM reasoning framework is the Codegen 2B-mono, initialized from Codegen 2B-multi and specifically trained on a corpus of Python code. Codegen 2B-multi is derived from Codegen 2B-nl and further trained on an extensive collection of code data from various programming languages. Codegen 2B-nl is randomly initialized and trained on the Pile, a vast English text corpus containing 825.18 million words.

The policy network in our LRM-MARL framework can be any cooperative MARL method. In our experiments, we selected representative methods from the two primary categories of value-based and policy-based approaches, such as QMIX[4], COMA[5], QTRAN[6], and value-decomposition networks (VDN)[7].

4.4 Better collaborative performance of our framework

We employ the ToM reasoner in our LRM-MARL partner modelling framework for better ToM reasoning and partner modelling in our proposed Reason tasks. As illustrated in Figs. 7 and 8, our framework outperforms multiple MARL baselines in terms of convergence speed, variance and average episode rewards, achieving better cooperation among agents. The complex map requires more sophisticated coordination due to increased elements and information, sparser rewards and an expanded search space. Table 1 compares the average episode rewards between the MARL baseline COMA and our ToM-enhanced COMA in the complex map. Our framework achieves more robust partner modelling, better cooperative performance, and faster convergence rates than the MARL baselines across different maps, indicating the

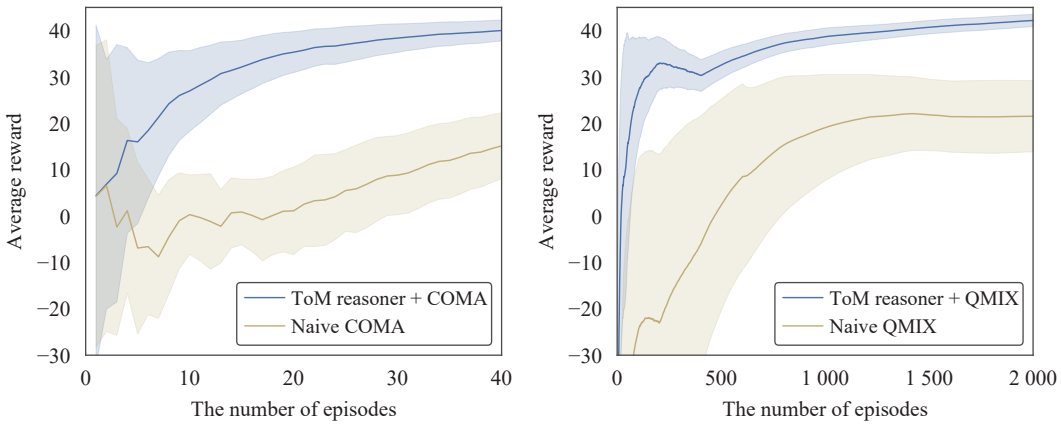

Fig. 7 Average episode reward comparison between several MARL baseline COMA (left), a policy-based method, QMIX (right), a value-based method, and our partner modelling framework. Our proposed ToM reasoner module further improves the performance of baseline methods, achieving faster convergence speed and smaller variances.

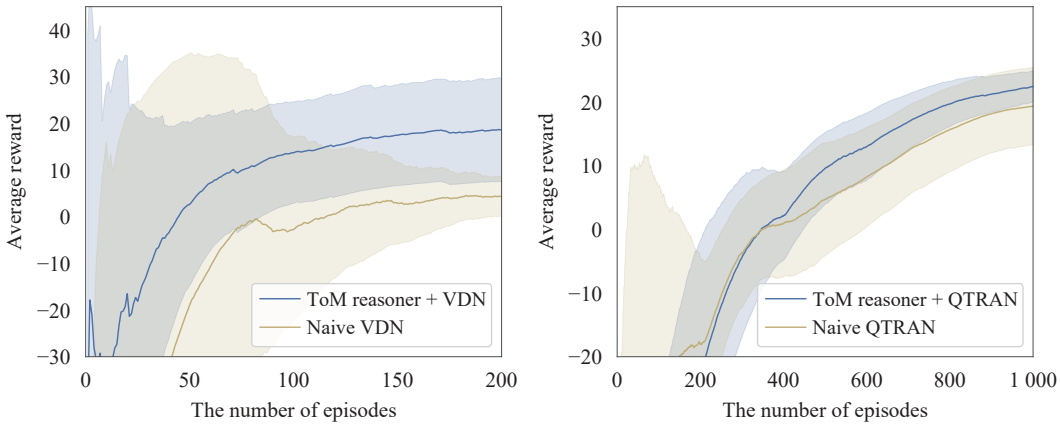

Fig. 8 Average episode reward comparison between several MARL baseline VDN (left), QTRAN (right), and our partner modelling framework. Our proposed ToM reasoner module further improves the performance of baseline methods, achieving faster convergence speed and smaller variances.

Table 1 The average episode reward results in a more challenging map between the cooperative MARL baseline COMA and our partner modelling framework.

Methods	Average reward	Variance
Naive COMA	26.86	±2.75
+ ToM reasoner	**32.86**	±**0.93**

generalizability of our framework in partner understanding and modelling.

The reasoning capabilities of our ToM reasoner are crucial for the success of our proposed framework, achieving more effective partner modelling and decision-making. By effectively comprehending the given task, narrowing the search space, and achieving beneficial ToM reasoning for partner belief, our partner modelling framework contributes to an accelerated training process and enhances cooperative reward performance. These experimental findings highlight the effectiveness of incorporating a brain-inspired partner modelling framework to enhance the ToM ability of traditional MARL agents. These experimental results suggest that the information extractor, the large reasoning model and the LAM module in our ToM reasoner contribute to such robust collaboration performance and efficiency.

4.5 Ablation study analysis

As shown in Fig. 9, we explore the effects of the LRM size in further experiments where LRMs at varying scales contain different numbers of neurons. Compared with the baselines, our proposed ToM reasoners with LRMs of different scales achieve notable performance enhancements and faster convergence compared to the baselines, demonstrating the adaptability of our framework and the effectiveness of our proposed ToM reasoner in simulating PFC structures. As the scale of the LRM increases, our framework can achieve better collaboration results and faster convergence. This result aligns with existing neuroimaging research[63, 64], which suggests that the larger PFC volume and greater PFC thickness are associated with stronger capacity, leading to better executive performance in decision-making tasks.

Many cognitive disorders of the human brain stem from a common factor: the disruption of neural activity within the PFC[56]. Cognitive disruption can sometimes arise from unexpected injuries, leading to changes in the size of the PFC. Recent biological research[65, 66] has focused on how these changes affect its functions. However, the minicolumnar basis of the PFC remains poorly understood due to technological constraints, presenting an open and challenging question for further analysis of the PFC[56]. Therefore, simulating the PFC may facilitate the development of new hypotheses and contribute to a more comprehensive understanding of neuroscience. Our previous analysis indicates that our ToM reasoner is biologically plausible and exhibits multiscale similarities with the PFC in the biological decision-making structure. Therefore, further computational experiments may provide computational insights into the study of the PFC at both the functional and structural scales.

In our further ablation analysis, we conduct extensive experiments to validate the effectiveness of the LAM module in our proposed ToM reasoner, as illustrated in Fig. 10. Our ablation study demonstrates that the LAM module effectively maps and transforms from the LRM semantic space to the MARL state-action space, facilitating efficient collaboration among agents and enhancing cooperative performance.

5 Discussions

Although our LRM-MARL framework achieves better partner modelling and cooperation performance, there are still limitations in various aspects. To simulate the PFC for robust and adaptive partner modelling, we developed a ToM reasoner within the LRM-MARL framework. However, since the LRM in our ToM reasoner is a purely

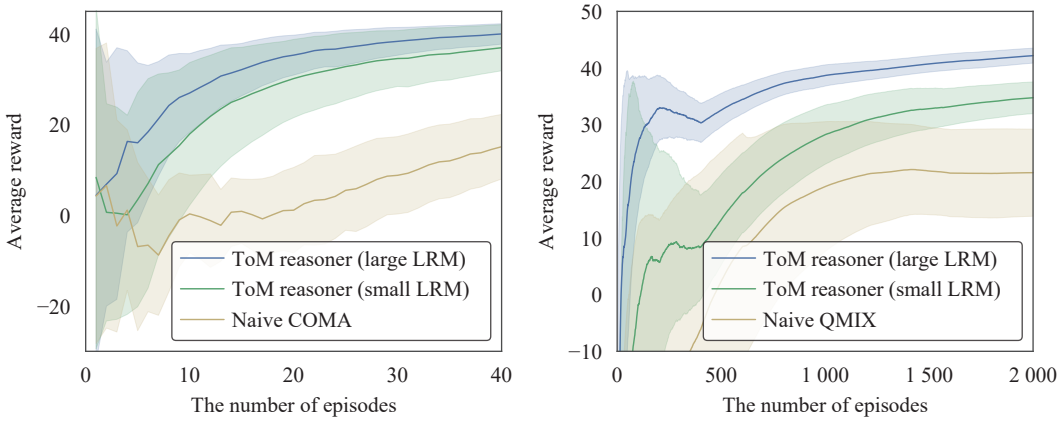

Fig. 9 Ablation results of different LRM scales for our ToM reasoner on COMA (left) and QMIX (right): Larger scales yield better performance and faster convergence. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

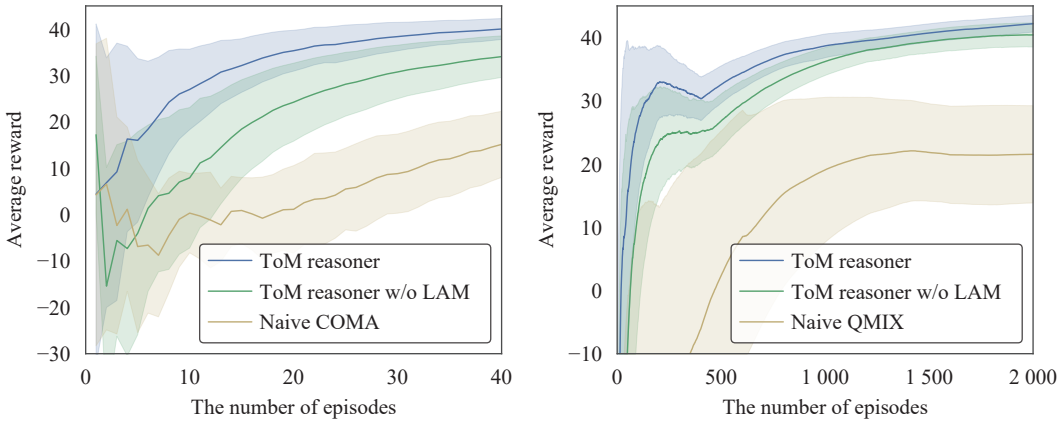

Fig. 10 Ablation results of the LAM module: Enhanced performance and accelerated convergence. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

text-based LRM, it lacks the ability to process multimodal information. Therefore, our framework requires the design of critical environmental features and prompt engineering in more complex environments to create more practical knowledge templates.

While our ToM reasoner demonstrates improved reasoning ability, it is challenging to adapt quickly to the variability of human behavior in complex environments. The incorporation of diverse trajectory data, such as diverse human trajectory data, into the LRM training process, could mitigate this limitation and improve its adaptability. Therefore, constructing a partner modelling framework with multimodal capabilities and faster adaptation to changing partner behaviors is a promising direction for future research.

6 Conclusions

Inspired by the biological decision-making structure and the reasoning process in the PFC, we have developed an LRM-improved MARL framework. This framework aims for robust and adaptive partner modelling with ToM reasoning capabilities in cooperative MARL. We incorporate our proposed framework into various competitive MARL methods, achieving better reward scores and sample efficiency. Further experimental analyses reveal the adaptivity, robustness and generalization of our proposed framework, indicating its potential applicability to more complex tasks. We also discuss the limitations of our LRM-MARL framework and propose possible solutions as future research directions.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, China (No. XDA27040200), the Beijing Nova Program, China (No. 20230484369), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences, China.

Declarations of conflict of interest

The authors declare that they have no conflicts of interest to this work.

References

[1] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. Mckinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, vol. 575, no. 7782, pp. 350–354, 2019. DOI: 10.1038/s41586-019-1724-z.

[2] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, Y. Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1787, 2022.

[3] L. Meng, M. Wen, C. Le, X. Li, D. Xing, W. Zhang, Y. Wen, H. Zhang, J. Wang, Y. Yang, B. Xu. Offline pre-trained multi-agent decision transformer. *Machine Intelligence Research*, vol. 20, no. 2, pp. 233–248, 2023. DOI: 10.1007/s11633-022-1383-7.

[4] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, S. Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, vol. 21, no. 1, Article number 178, 2020.

[5] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, S. Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, USA, pp. 2974–2982, 2018. DOI: 10.1609/aaai.v32i1.11794.

[6] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, Y. Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Pro-*

ceedings of the 36th International Conference on Machine Learning, Long Beach, USA, pp. 5887–5896, 2019.

[7] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, T. Graepel. Value-decomposition networks for cooperative multi-agent learning, [Online], Available: https://arxiv.org/abs/1706.05296, 2017.

[8] S. V. Albrecht, P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, vol. 258, pp. 66–95, 2018. DOI: 10.1016/j.artint.2018.01.002.

[9] Z. Tian, Y. Wen, Z. Gong, F. Punakkath, S. Zou, J. Wang. A regularized opponent model with maximum entropy objective. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 602–608, 2019. DOI: 10.24963/ijcai.2019/85.

[10] H. He, J. Boyd-Graber, K. Kwok, H. Daumé III. Opponent modeling in deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, New York City, USA, pp. 1804–1813, 2016.

[11] Y. Wen, Y. Yang, R. Luo, J. Wang, W. Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[12] X. Li, Z. Ni, J. Ruan, L. Meng, J. Shi, T. Zhang, B. Xu. Mixture of personality improved spiking actor network for efficient multi-agent cooperation, [Online], Available: https://arxiv.org/abs/2305.05898, 2023.

[13] A. Shih, A. Sawhney, J. Kondic, S. Ermon, D. Sadigh. On the critical role of conventions in adaptive human-AI collaboration. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[14] OpenAI. GPT-4 technical report, [Online], Available: https://arxiv.org/abs/2303.08774, 2023.

[15] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, D. Yang. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 1339–1384, 2023. DOI: 10.18653/v1/2023.emnlp-main.85.

[16] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4, [Online], Available: https://arxiv.org/abs/2303.12712, 2023.

[17] R. Anil et al. PaLM 2 technical report, [Online], Available: https://arxiv.org/abs/2305.10403, 2023.

[18] D. A. Boiko, R. MacKnight, B. Kline, G. Gomes. Autonomous chemical research with large language models. *Nature*, vol. 624, no. 7992, pp. 570–578, 2023. DOI: 10.1038/s41586-023-06792-0.

[19] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, W. Zaremba. Evaluating large language models trained on code, [Online], Available: https://arxiv.org/abs/2107.03374, 2021.

[20] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, C. Xiong. CodeGen: An open large language model for code with multi-turn program synthesis. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[21] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, J. Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, vol. 18, no. 6, Article number 186345, 2024. DOI: 10.1007/s11704-024-40231-1.

[22] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 377, 2023.

[23] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[24] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, A. Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, vol. 2024, 2024.

[25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 159, 2020.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1800, 2022.

[27] N. Feltovich. Reinforcement-based VS. belief-based learning models in experimental asymmetric-information games. *Econometrica*, vol. 68, no. 3, pp. 605–641, 2000. DOI: 10.1111/1468-0262.00125.

[28] D. Von Winterfeldt. Bridging the gap between science and decision making. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. S3, pp. 14055–14061, 2013. DOI: 10.1073/pnas.1213532110.

[29] H. Seo, D. Lee. Neural basis of learning and preference during social decision-making. *Current Opinion in Neurobiology*, vol. 22, no. 6, pp. 990–995, 2012. DOI: 10.1016/j.conb.2012.05.010.

[30] H. L. Gallagher, C. D. Frith. Functional imaging of "theory of mind". *Trends in cognitive sciences*, vol. 7, no. 2, pp. 77–83, 2003. DOI: 10.1016/S1364-6613(02)00025-6.

[31] C. Frith, U. Frith. Theory of mind. *Current Biology*, vol. 15, no. 17, pp. R644–R646, 2005. DOI: 10.1016/j.cub.2005.08.041.

[32] A. M. Leslie. Pretense and representation: The origins of "theory of mind". *Psychological Review*, vol. 94, no. 4, pp. 412–426, 1987. DOI: 10.1037/0033-295X.94.4.412.

[33] C. E. V. Mahy, L. J. Moses, J. H. Pfeifer. How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience*, vol. 9, pp. 68–81, 2014. DOI: 10.1016/j.dcn.2014.01.002.

[34] L. Yuan, X. Gao, Z. Zheng, M. Edmonds, Y. N. Wu, F. Rossano, H. Lu, Y. Zhu, S. C. Zhu. In situ bidirectional human-robot value alignment. *Science Robotics*, vol. 7, no. 68, Article number 4183, 2022. DOI: 10.1126/scirobotics.abm4183.

[35] Y. Wang, F. Zhong, J. Xu, Y. Wang. ToM2C: Target-oriented multi-agent communication and cooperation with theory of mind. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[36] X. Li, T. Zhang, C. Liu, L. Meng, B. Xu. Long short-term reasoning network with theory of mind for efficient multi-agent cooperation. In *Proceedings of International Joint Conference on Neural Networks*, Yokohama, Japan, 2024. DOI: 10.1109/IJCNN60899.2024.10650244.

[37] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLoS One*, vol. 12, no. 4, Article number e0172395, 2017. DOI: 10.1371/journal.pone.0172395.

[38] Z. Wu, K. Li, H. Xu, Y. Zang, B. An, J. Xing. L2E: Learning to exploit your opponent. In *Proceedings of International Joint Conference on Neural Networks*, Padua, Italy, 2022. DOI: 10.1109/IJCNN55064.2022.9892077.

[39] K. Valmeekam, A. Olmo, S. Sreedharan, S. Kambhampati. Large language models still can′t plan (A benchmark for LLMs on planning and reasoning about change), [Online], Available: https://arxiv.org/abs/2206.10498, 2022.

[40] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, M. Zhou. CodeBERT: A pre-trained model for programming and natural languages. In *Proceedings of Findings of the Association for Computational Linguistics*, pp. 1536–1547, 2020. DOI: 10.18653/v1/2020.findings-emnlp.139.

[41] M. Sclar, S. Kumar, P. West, A. Suhr, Y. Choi, Y. Tsvetkov. Minding language models′ (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, pp. 13960–13980, 2023. DOI: 10.18653/v1/2023.acl-long.780.

[42] D. M. Amodio, C. D. Frith. Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, vol. 7, no. 4, pp. 268–277, 2006. DOI: 10.1038/nrn1884.

[43] H. D. Schlinger. Theory of mind: An overview and behavioral perspective. *The Psychological Record*, vol. 59, no. 3, pp. 435–448, 2009. DOI: 10.1007/BF03395673.

[44] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. Ali Eslami, M. Botvinick. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 4218–4227, 2018.

[45] A. Tabrez, M. B. Luebbers, B. Hayes. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, vol. 1, no. 4, pp. 259–267, 2020. DOI: 10.1007/s43154-020-00019-0.

[46] K. Daniel. *Thinking, Fast and Slow*, USA: Farrar, Straus and Giroux, 2013.

[47] P. Croskerry, D. A. Petrie, J. B. Reilly, G. Tait. Deciding about fast and slow decisions. *Academic Medicine*, vol. 89, no. 2, pp. 197–200, 2014. DOI: 10.1097/ACM.0000000000000121.

[48] R. L. E. P. Reniers, R. Corcoran, B. A. Völlm, A. Mashru, R. Howard, P. F. Liddle. Moral decision-making, tom, empathy and the default mode network. *Biological Psychology*, vol. 90, no. 3, pp. 202–210, 2012. DOI: 10.1016/j.biopsycho.2012.03.009.

[49] H. F. Kim, O. Hikosaka. Parallel basal ganglia circuits for voluntary and automatic behaviour to reach rewards. *Brain*, vol. 138, no. 7, pp. 1776–1800, 2015. DOI: 10.1093/brain/awv134.

[50] M. Jahanshahi, I. Obeso, J. C. Rothwell, J. A. Obeso. A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. *Nature Reviews Neuroscience*, vol. 16, no. 12, pp. 719–732, 2015. DOI: 10.1038/nrn4038.

[51] D. Badre, M. D′Esposito. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, vol. 10, no. 9, pp. 659–669, 2009. DOI: 10.1038/nrn2667.

[52] E. K. Miller, J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, vol. 24, pp. 167–202, 2001. DOI: 10.1146/annurev.neuro.24.1.167.

[53] J. Rowe, K. Friston, R. Frackowiak, R. Passingham. Attention to action: Specific modulation of corticocortical interactions in humans. *NeuroImage*, vol. 17, no. 2, pp. 988–998, 2002. DOI: 10.1006/nimg.2002.1156.

[54] M. Wang, Y. Yang, C. J. Wang, N. J. Gamo, L. E. Jin, J. A. Mazer, J. H. Morrison, X. J. Wang, A. F. T. Arnsten. NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron*, vol. 77, no. 4, pp. 736–749, 2013. DOI: 10.1016/j.neuron.2012.12.032.

[55] P. Brown, C. Marsden. What do the basal ganglia do?. *The Lancet*, vol. 351, no. 9118, pp. 1801–1804, 1998. DOI: 10.1016/S0140-6736(97)11225-9.

[56] I. Opris, R. E. Hampson, G. A. Gerhardt, T. W. Berger, S. A. Deadwyler. Columnar processing in primate pFC: Evidence for executive control microcircuits. *Journal of cognitive neuroscience*, vol. 24, no. 12, pp. 2334–2347, 2012. DOI: 10.1162/jocn_a_00307.

[57] M. Donoso, A. G. E. Collins, E. Koechlin. Foundations of human reasoning in the prefrontal cortex. *Science*, vol. 344, no. 6191, pp. 1481–1486, 2014. DOI: 10.1126/science.1252254.

[58] S. A. Bunge, C. Wendelken, D. Badre, A. D. Wagner.

Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, vol. 15, no. 3, pp. 239–249, 2005. DOI: 10.1093/cercor/bhh126.

[59] M. Kosinski. Theory of mind may have spontaneously emerged in large language models, [Online], Available: https://arxiv.org/abs/2302.02083, 2023.

[60] H. M. Wellman, D. Cross, J. Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, vol. 72, no. 3, pp. 655–684, 2001. DOI: 10.1111/1467-8624.00304.

[61] T. Korkiakangas, K. Dindar, A. Laitila, E. Kärnä. The Sally–Anne test: An interactional analysis of a dyadic assessment. *International Journal of Language & Communication Disorders*, vol. 51, no. 6, pp. 685–702, 2016. DOI: 10.1111/1460-6984.12240.

[62] M. Sommer, K. Döhnel, B. Sodian, J. Meinhardt, C. Thoermer, G. Hajak. Neural correlates of true and false belief reasoning. *NeuroImage*, vol. 35, no. 3, pp. 1378–1384, 2007. DOI: 10.1016/j.neuroimage.2007.01.042.

[63] A. Maat, N. E. M. van Haren, C. F. Bartholomeusz, R. S. Kahn, W. Cahn. Emotion recognition and theory of mind are related to gray matter volume of the prefrontal cortex in schizophrenia. *European Neuropsychopharmacology*, vol. 26, no. 2, pp. 255–264, 2016. DOI: 10.1016/j.euroneuro.2015.12.013.

[64] P. Yuan, N. Raz. Prefrontal cortex and executive functions in healthy adults: A meta-analysis of structural neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, vol. 42, pp. 180–192, 2014. DOI: 10.1016/j.neubiorev.2014.02.005.

[65] A. F. T. Arnsten. Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience*, vol. 10, no. 6, pp. 410–422, 2009. DOI: 10.1038/nrn2648.

[66] S. Alvarado, M. Tajerian, M. Millecamps, M. Suderman, L. S. Stone, M. Szyf. Peripheral nerve injury is accompanied by chronic transcriptome-wide changes in the mouse prefrontal cortex. *Molecular Pain*, vol. 9, Article number 21, 2013. DOI: 10.1186/1744-8069-9-21.

Xiyun Li received the B. Sc. degree in computer science and technology from Lanzhou University, China in 2020. He is currently a Ph. D. degree candidate in pattern recognition and intelligent system at both the Institute of Automation, Chinese Academy of Sciences, China, and the University of Chinese Academy of Sciences, China.

His research interests include reinforcement learning, brain-inspired cognitive models and speech separation.

E-mail: lixiyun2020@ia.ac.cn

ORCID iD: 0009-0001-9517-7030

Tielin Zhang received the Ph. D. degree in brain-inspired intelligence from the Institute of Automation, Chinese Academy of Sciences, China in 2016. He is with the Center for Excellence in Brain Science and Intelligence Technology, State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Neuroscience, Chinese Academy of Sciences, China, and also with School of Artificial Intelligence, University of Chinese Academy of Sciences, China.

His research interests include theoretical research on neural dynamics and spiking neural networks.

E-mail: zhangtielin@ion.ac.cn (Corresponding author)

ORCID iD: 0000-0002-5111-9891

Chenghao Liu received the Ph. D. degree in neuroscience from Brandeis University, USA in 2021. He is a postdoctoral researcher at the The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China.

His research interests include computational modeling of nervous systems and neural networks.

E-mail: chenghao.liu@ia.ac.cn

Shuang Xu received the B. Sc. and the M. Sc. degrees in measuring and testing technologies and instruments from Yanshan University, China in 2001 and 2004, respectively, and the Ph. D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, China in 2009. She is a professor at the Institute of Automation, Chinese Academy of Sciences, China.

Her research interests include natural language processing and understanding, and human-AI hybrid intelligence.

E-mail: shuang.xu@ia.ac.cn

Bo Xu received the B. Sc. degree in electrical engineering from Zhejiang University, China in 1988, and the M. Sc. and the Ph. D. degrees in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, China in 1992 and 1997, respectively. He is a professor, the director of the Institute of Automation, Chinese Academy of Sciences, and also the deputy director of the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, China.

His research interests include brain-inspired intelligence, brain-inspired cognitive models, and natural language processing and understanding.

E-mail: xubo@ia.ac.cn (Corresponding author)

ORCID ID: 0000-0002-1111-1529