



Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning

Bidipta Sarkar 

Stanford University

Stanford, United States of America


bidiptas@cs.stanford.edu

C. Karen Liu 

Stanford University

Stanford, United States of America

karenliu@cs.stanford.edu

Warren Xia 

Stanford University

Stanford, United States of America

waxia@cs.stanford.edu

Dorsa Sadigh 

Stanford University

Stanford, United States of America

dorsa@cs.stanford.edu

ABSTRACT

Communicating in natural language is a powerful tool in multi-agent settings, as it enables independent agents to share information in partially observable settings and allows zero-shot coordination with humans. However, most prior works are limited as they either rely on training with large amounts of human demonstrations or lack the ability to generate natural and useful communication strategies. In this work, we train language models to have productive discussions about their environment in natural language without any human demonstrations. We decompose the communication problem into *listening* and *speaking*. Our key idea is to leverage the agent’s goal to *predict useful information about the world* as a dense reward signal that guides communication. Specifically, we improve a model’s listening skills by training them to predict information about the environment based on discussions, and we simultaneously improve a model’s speaking skills with multi-agent reinforcement learning by rewarding messages based on their influence on other agents. To investigate the role and necessity of communication in complex social settings, we study an embodied social deduction game based on AMONG US, where the key question to answer is the identity of an adversarial imposter. We analyze emergent behaviors due to our technique, such as accusing suspects and providing evidence, and find that it enables strong discussions, doubling the win rates compared to standard RL. We release our code and models at <https://socialdeductionllm.github.io/>.





CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**; *Stochastic games*; *Cooperation and coordination*; • **Information systems** → **Language models**.

KEYWORDS

Language Models; Multi-Agent Reinforcement Learning; Social Deduction; LLM Agents

ACM Reference Format:

Bidipta Sarkar , Warren Xia , C. Karen Liu , and Dorsa Sadigh . 2025. Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 INTRODUCTION

A longstanding goal of multi-agent artificial intelligence is the development of independent agents that can communicate using a shared language. Communication is especially necessary in “partially observable” settings, where each agent only has a limited view of the world and therefore benefits from sharing knowledge with other agents to achieve its goal. In particular, “social deduction” games are settings where each agent’s goal is to deduce information about the environment by communicating with other agents – requiring each player to learn how to parse messages from other players while effectively sharing important information needed for game completion.

In this work, we study the hidden-role game of AMONG US [18] as a specific instance of a challenging social deduction game to investigate the importance of communication, illustrated in Fig. 1. Hidden-role games [4, 19] are a class of environments where players are split into an uninformed majority and a smaller informed hidden subteam, which we refer to as *crewmates* and *imposters* respectively. These two teams are adversaries, resulting in a zero-sum game, where the goal of the crewmates is to deduce the identity of imposters to vote them out. Unlike other popular hidden role games such as the game of Mafia [2], where statements from players are unfalsifiable, AMONG US is based in a 2D embodied environment, allowing discussions and intuitions to be grounded in specific observations. In the game, crewmates try to complete an assigned set of tasks scattered across the environment while imposters try to kill all crewmates. If a player reports the corpse of an eliminated crewmate – killed by an imposter – the game moves to a discussion phase with a free-form chat followed by a voting period, where all players vote to eject a suspected imposter. For crewmates, success in the discussion phase would mean correctly voting out the imposter, while success for imposters means avoiding suspicion from the crewmates to continue staying in the game as long as possible. This highlights the importance of communication during the discussion phase as crewmates need to learn to effectively utilize



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

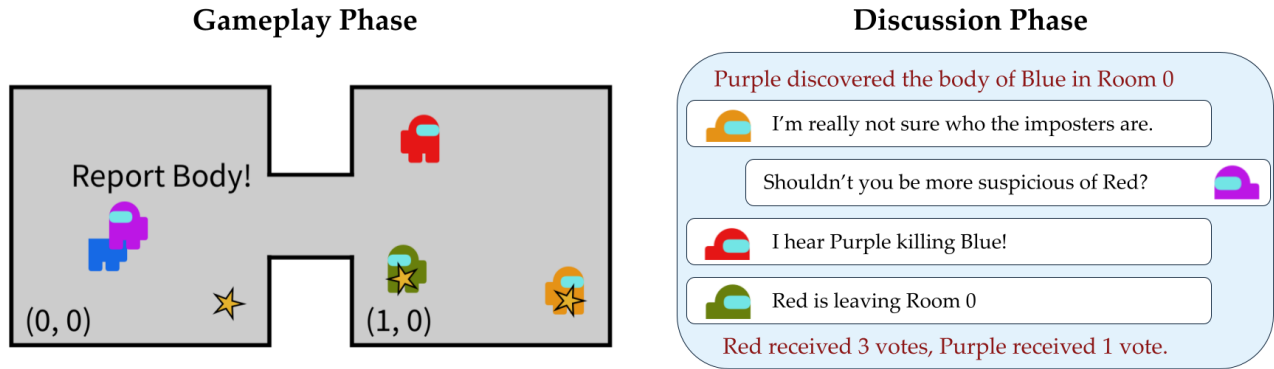


Figure 1: Examples of the gameplay and discussion phases of AMONG Us. During gameplay, all agents navigate a 2D grid environment (a 1-by-2 grid in this case, with two rooms at (0,0) and (1,0)), where agents can see everything in their same room. Here, the red, green, and yellow agents are in room (1,0), and the purple and blue agents are in room (0,0). Crewmates can perform tasks (indicated by the stars – in this example there are 3 tasks), while imposters kill crewmates. Here, the orange and green agents are working on tasks. Agents can also report dead bodies, as the purple agent is currently doing, which initiates the discussion phase. During discussion phases, agents leverage large language models to generate free-form messages guided by our framework encouraging effective speaking and listening within the crewmates and finally vote out a suspected imposter. The example discussion shown on the right is based on a generated discussion from our trained models.

the discussion phase to vote out imposters in an adversarial setting. For the rest of this paper, we study the game of AMONG Us from the perspective of crewmates attempting to perform tasks, identify imposters, and win the game.

In multi-agent environments, an effective technique for training strong cooperative and competitive agents is multi-agent reinforcement learning (MARL), which enables artificial agents to achieve superhuman levels of performance in competitive games such as StarCraft [36], and cooperative games such as Overcooked [5, 31] and Hanabi [13]. However, in settings where communication in natural language is necessary, existing MARL techniques often struggle as they require large datasets of task-specific human communication data to perform on-par with humans [8]. This fundamentally limits the agents’ ability to communicate at human-level and is not practical for learning in settings where these datasets do not readily exist. The game of AMONG Us falls into this category, where communication is necessary to reason and progress in the game. Therefore, we would like to find an approach that learns to communicate effectively and convincingly without requiring large amounts of task-specific human data. However, the major challenge in learning to communicate without access to large amounts of human data is that novice agents do not have a strong signal for understanding the helpfulness of the messages they send (*speaking*) or for learning the meaning of messages from other players (*listening*). In particular, the sparse reward signal the agents receive when winning the game is not informative enough to reinforce high-quality discussions between agents. Our key insight is that we can leverage the agents’ instrumental goal of *predicting useful information about the world* – e.g., the identity of imposters – as a dense reward to provide a higher-quality signal that can enable more informative communication during the discussion phase and potentially higher performance policies.

We propose an approach that rewards a message generated during the discussion phase based on how the other crewmates’ beliefs on the identity of the imposter changes. Each crewmate wants to send messages that help other crewmates be more certain about the true identity of the imposter. However, this only explains how to learn to “speak” assuming that the other agents can appropriately update their belief about the world given a message. We also need to ensure the agents know how to “listen” and update beliefs appropriately. To encourage this, we additionally add an *imposter prediction* signal to guide the agent’s learning to predict the true identity of the imposter after each message. By training agents to speak and listen effectively, we enable the agents to self-improve their discussion abilities. Further, to encourage *listening* and *speaking* in natural language during the discussion phase of the game, we tap into the power of large language models (LLMs), unspecialized models trained with large amounts of human language data. Specifically, we initialize crewmates as LLMs capable of communicating via natural language. Recent advances in foundation models have demonstrated some reasoning abilities [3, 27], including understanding social scenarios [20], but even the strongest language models today are weak at self-critiquing [34] or performing theory of mind reasoning [33], limiting their ability to improve their listening skills based on their own feedback. However, by training LLMs within our proposed framework of encouraging listening and speaking with auxiliary dense rewards for helping other crewmates vote out the correct imposter, we overcome this limitation, enabling the self-improvement of these models over time.

To evaluate our framework, we analyze the success rate of crewmates against both pretrained and adaptive imposters, and find that crewmates form a robust communication strategy. We find that our technique results in emergent behavior commonly found in real games of AMONG Us between humans, such as directly accusing

players and providing evidence to help other crewmates [22]. We also find that our augmentation to discussions results in two times higher success rates relative to standard RL along with over three times higher success rates relative to base models that are over four times larger than our models, highlighting the importance of our discussion strategies.

2 RELATED WORK

In this section, we review related work on emergent communication, prior works that use language models as agents in embodied settings, and past works integrating language models with RL.

Emergent Communication. A major topic in MARL is emergent communication between agents, especially in the context of reference games and repeated reference games, where a speaker knows the ground-truth answer to a question (e.g., a specific image out of a set of images that needs to be referred to). Then, the speaker needs to communicate to the listener, who later needs to choose the item being referenced either over one or repeated interactions. Prior work has shown that humans tend to quickly adapt to such tasks [26], naturally using theory of mind reasoning to determine the intents of speakers [9]. Further, Hawkins et al. [12] showed that language models can also learn to adapt to human conventions via continual learning. Without using human natural language data, Lazaridou et al. [23] and Havrylov and Titov [11] use symbolic cheap-talk signals to solve referential games. Our framework of social deduction games, however, is more challenging as each agent does not know the ground truth answer, so teams must communicate to collectively learn the answer. Therefore, our domain does not have as clear of a distinction between “speakers” who have knowledge and “listeners” who need to gain answers as agents in social deduction games must play both roles.

Language Models Agents. A large body of prior work use LLMs’ access to internet scale data for task planning and decision making. In robotics, prior works explore how language models can be used to plan out a sequence of high-level primitives given an instruction in natural language [1, 17, 24]. In a virtual gaming setting, Park et al. [29] uses ChatGPT to simulate members of a small virtual town. Although there is no specific task or mechanism for “training” these agents, they demonstrate the use of a long-term memory stream to store memories beyond the context length of the language models, enabling the formation of social networks. This technique of having external memory has later been used to learn “skills” in a single-player environment [38] and for coordination in multi-agent environments [10]. These works demonstrate that language models are capable of controlling agents in a wide range of settings, which is key to our motivation to directly use language models as a strong starting point for agents operating in more challenging environments such as social deduction games.

Reinforcement Learning with Foundation Models. Some works also combine language models with reinforcement learning. Cicero [8] is an AI for the game of Diplomacy that uses a dialogue-conditional action model from human actions and trains a dialogue-free model using RL to choose actions. Cicero uses an “intent” embedding to connect the dialogue generation and strategic reasoning components. This allows Cicero to communicate with other agents in a way that feels natural to other players, but it prevents the RL

model from directly controlling the generated messages, potentially limiting improvements in message quality. Another drawback is that this technique requires a large number of human demonstrations, which may be impractical in many settings.

Foundation models have been effective in both providing rewards and as a base model for policies. Hu and Sadigh [14] and Kwon et al. [21] use language models as reward signals to train a separate network to follow a specific coordination strategy. We similarly use the LLM to provide denser rewards during the discussion phase, but we train the LLM itself instead of a separate policy.

Outside of the embodied setting, reinforcement learning has also been key to improving the chat capabilities of LLMs. Ouyang et al. [28] demonstrates the effectiveness of reinforcement learning from human feedback (RLHF), where a reward model is trained using human feedback and an LLM is fine-tuned using a modification of the PPO algorithm to improve its performance. Yuan et al. [39] extends this by allowing the LLM to be its own reward model and generate its own data for self-improvement, similar to how we use the LLM’s own change in beliefs as a reward signal. However, a crucial difference is that our reward model remains grounded in an environment by design due to the imposter prediction training signal. This means that we do not need to rely on the ability of pretrained LLMs to critique their own generations, enabling us to use smaller language models and correct logical errors over time.

3 PRELIMINARIES

We model social deduction games, such as AMONG US, as a variant of the partially observable Markov game (POMG) [25] that includes a question whose answer must be deduced through interacting with players and the rest of the environment. Our modified POMG can be described by a tuple $(n, \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathcal{O}, \gamma, \mathcal{Q}, q)$, where n is the number of players, \mathcal{S} is the joint (hidden) state space and \mathcal{A} is the joint action space. The transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, is the probability of reaching a state given the current state and joint action. The reward function $r : \mathcal{S} \rightarrow \mathbb{R}^n$, gives a real value reward for each state transition to each player, and γ is the reward discount. The observation function, $\mathcal{O} : \mathcal{S} \rightarrow \mathcal{O}^n$, generates the player-specific observations from the state.

Our POMG has additional terms for the task of social deduction, which are the set of all possible answers to the deduction problem $\mathcal{Q} \subseteq \mathcal{A}$ and the correct answer $q \in \mathcal{Q}$. In social deduction games, agents will be given opportunities to answer the question as a literal action (i.e. voting in AMONG US or choosing the correct object in reference games), and at those steps the correct action to take is q .

The trajectory up to time t is defined as a sequence of joint observations and actions: $\tau_t = (o_0, a_0, \dots, a_{t-1}, o_t)$. An individual player only experiences their own action-observation history (AOH), which is defined for player i as $\tau_t^i = (o_0^i, a_0^i, \dots, a_{t-1}^i, o_t^i)$, and they follow a stochastic policy $\pi^i(a^i | \tau^i)$. In the game of AMONG US, the AOH consists of past observations supplied by the environment, past embodied actions, and all prior discussions with the other players.

Language Models. Language models are trained to model the probability of a sequence of discrete tokens, where each token represents a string of natural language. For a sequence of tokens,

$W = \{w_0, w_1, \dots, w_k\}$, the probability of the sequence being generated is $p(W) = \prod_{j=0}^k p(w_j | w_{<j})$, so causal language models predict the distribution of the next token conditioned on all prior tokens.

Our AMONG Us environment is designed such that each observation at time step t is a sequence of tokens $o_t = W^t = \{w_0, w_1, \dots, w_k\}$ and each action at time step t is a single token $a_t = w_t$, allowing us to use language models as the policy for each agent. The AOH is a sequence of tokens, so language models can sample the next action by predicting the next token in the sequence, constrained to the set of legal actions at that timestep.

In this work, we use the RWKV language model [30], a recurrent language model based on a linear attention mechanism, as the pre-trained foundation model. We choose RWKV over more common transformer-based models [35], because the recurrent formulation allows us to generate RL trajectories with a constant time and space complexity per token, and RWKV enables unbounded-context training using truncated backpropagation through time. This is especially important since AMONG Us trajectories often reach tens of thousands of tokens in length per player, which would require significantly more compute for classic attention-based models. Empirically, RWKV has also performed on-par with transformer-based models, especially in decision-making tasks [7] and long-context understanding [15], making it the ideal choice for this study.

4 THE GAME OF AMONG US

In this section, we describe the key design decisions of our implementation of the hidden-role game of AMONG Us. Our goal is to create an environment where agents can ground their discussion based on evidence in the environment. A more complete description of the game is in Appendix A.

Role Assignment. At the start of the game, each player is either assigned as an *imposter* or a *crewmate*. The crewmates are not informed of the identities of the other players, but all imposters are informed of the identities of the other players at the start.

In our setting, we assign one player to be the imposter and the other $n - 1$ players as crewmates. The crewmates are assigned a set of N tasks, scattered across the environment. As an example, $N = 3$ in the example in Fig. 1.

Gameplay Phase. During the gameplay phase, players simultaneously move in an embodied environment, receiving observations from the environment and taking actions, as illustrated in Fig. 2. Players freely move around a $W \times H$ grid of rooms during the gameplay phase, receiving new observations o_t at each time step. All agents can move between adjacent rooms by choosing $a_{go\ to\ x}$, where x is a cardinal direction, or they can simply wait in the room by choosing a_{wait} . Crewmates can complete tasks in their current room by choosing a_{task} , but they are unable to observe the environment for N_{task_time} time steps, i.e., they will not be able to observe if a crewmate is being killed by an imposter while performing a task. Note that tasks are indistinguishable from one another, so we do not have different actions for different tasks. Imposters can kill crewmates by choosing $a_{kill,j}$ where j is a crewmate in the same room as them, but they have to wait $N_{cooldown}$ time steps between killing crewmates. Finally, crewmates can report dead bodies in their room by choosing $a_{report,j}$, where j is the corpse of player j , which initiates the discussion phase.

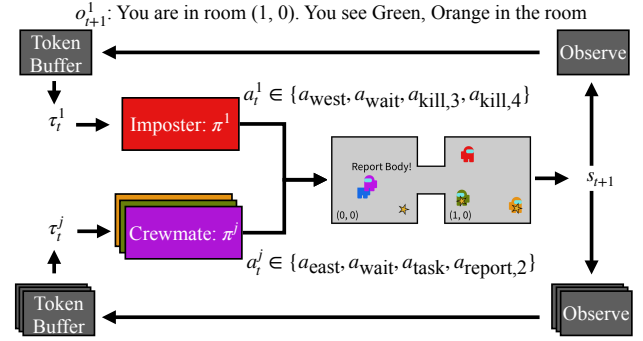


Figure 2: Diagram of the embodied gameplay loop. The environment starts by sending observations to all agents simultaneously and collects tokenized actions from a set of valid actions at each timestep.

The set of all valid actions are $a_{go\ to\ x}$, a_{task} , $a_{kill,j}$, $a_{report,j}$, and a_{wait} , where x is a cardinal direction and j is the name of a crewmate. The environment provides each player with the subset of actions that are valid at each timestep.

Discussion Phase. During the discussion phase, we cycle over each player twice in a random order and allow them to say a sentence, a_{talk} , in natural language. After this discussion, a voting phase begins, where each player votes for one player, k , they want to eject by choosing action $a_{vote,k}$. The player who gets the plurality of votes is ejected. If the imposter is not ejected, the game continues to the next gameplay phase, where crewmates can continue finishing tasks.

Before the discussion starts and between each discussion message, the environment also surveys each crewmate by asking who they would vote for, i.e., by querying them to pick an $a_{vote,k}$ as if they had to vote immediately. This action has no impact on the POMG, but it will be relevant for our training algorithm.

Note that the set of all voting actions is equal to the set of all possible “answers” in the social deduction game (Q), and voting out the correct imposter corresponds to q , the correct answer to the social deduction question.

Reward Structure. AMONG Us is fundamentally a team zero-sum game, so reward is based on whether crewmates or imposters win. If all tasks are completed or the imposter is ejected, the crewmates win with a reward of +1. However, if the number of imposters is ever greater than or equal to the number of crewmates, the imposters win, resulting in a crewmate reward of -1.

5 TRAINING LLM CREWMATES IN AMONG US

By defining an environment that only interfaces with players through natural language, we can directly use a language model as the policy $\pi^i(a^i | \tau^i)$ of an agent i . The action-observation histories of our agents τ^i are just strings of natural language, and new observations and actions can simply be appended to the end of the strings. Furthermore, when taking actions a^i , the outputs of the language model can be constrained to be one of the legal actions provided by the environment at each timestep. Following this procedure,

we construct an agent using a pretrained RWKV language model, which we define as policy π_{RWKV} .

Although the environment is designed to interface nicely with language models, we find that π_{RWKV} struggles to reason as crewmates in a zero-shot fashion in AMONG US, with models frequently voting to eject the wrong players. In this section, we describe our procedure for improving the performance of crewmates by enabling them to self-critique and use these scores to improve dialogue generation.

The first two subsections describe how to improve the performance of an individual learning crewmate, first describing a reinforcement learning procedure and then describing how to enhance communication by learning to listen and speak. The third subsection describes how to train the team of crewmates to be robust to adaptive imposters and different policies within the crewmate population.

5.1 Reinforcement Learning in AMONG US

To train a language model to take more effective actions without expert demonstrations, we can turn to reinforcement learning. Since AMONG US already provides rewards for winning, we can directly optimize this to produce a model π_{RL} that minimizes the following loss:

$$L_{\text{RL}}(\pi) = - \mathbb{E}_{\tau^i \sim \Pi} \sum_t \left[\gamma^t r_t^i + \lambda_{\text{NL}} \log \left(\frac{\pi(a_t^i | \tau_t^i)}{\pi_{\text{RWKV}}(a_t^i | \tau_t^i)} \right) \right], \quad (1)$$

where Π represents the joint policy that has π controlling agent i , and λ_{NL} is a hyperparameter controlling the strength of a soft KL constraint regularizing trained models to the base LLM to prevent discussions from moving out of natural language [28]. Note that the only reward signal is the sparse reward received at the end of the game along with additional rewards for completing tasks. In particular, there is very little signal for the effectiveness of its messages during discussions, which makes utilizing communication very difficult with just RL in practice. This sparse signal also makes identifying the imposter difficult in the multi-agent setting, because voting correctly may still result in a loss and voting incorrectly could result in a win if a plurality of agents vote for the imposter.

5.2 Enhancing Communication of Crewmates

To improve beyond the RL baseline, we can take advantage of the social deduction component of the game. In particular, each agent’s belief in choosing the correct answer $q \in Q$ will provide a stronger signal for learning the core components of the game and the means of communication relative to the RL baseline.

In this subsection, we discuss the key contributions of this work. Specifically, we highlight new loss terms to enhance both listening and speaking abilities, enabling crewmates to better utilize discussions.

Listening: Imposter Prediction. Suppose an agent is learning in a environment where it is partnered with expert crewmates who already know how to discuss the game. How can this agent learn to understand the meanings of environment observations and messages from other crewmates?

To effectively discuss the game of AMONG US, crewmates need to understand who the imposter is given their past observations

and the past messages. This prediction task can act as an auxiliary task that can guide the discussion phase to be more grounded and meaningful.

We directly train crewmates to improve their reasoning over imposters using the environment’s ground truth answer for the identity of the imposter. Specifically, we use the timesteps when the environment directly surveys the players for their beliefs over the imposters, which occurs between discussion messages, as the training signal. Note that this training signal does not specifically require human demonstration data; agents can learn to understand observations and messages from other players using any rollout buffer.

For every living crewmate, if they are asked to provide their beliefs regarding the identity of the imposter at timestep t , the *listening loss* for that timestep is

$$L_L(\pi, \tau_t^i) = -\log \pi(q | \tau_t^i), \quad (2)$$

where $q = a_{\text{vote},j}$ is the action representing choosing the correct imposter j , and τ_t^i is the AOH until timestep t , which may include prior discussions.

At the very start of the discussion phase, agents need to reflect on the probabilities of other agents being the imposter based on their observations during the gameplay phase. For instance, if a crewmate directly witnesses a murder, they should be very certain that the murderer is the imposter; our listening loss uses this signal to increase their certainty over the imposter.

By framing the task of identifying imposters using messages and observations as a supervised learning problem, agents learn to understand the meaning of messages, enabling them to vote out the correct imposter. Using this loss term, we can define two new policies. We can directly incorporate the listening loss into the RL policy, giving us the policy $\pi_{\text{RL}+L}$ that optimizes

$$L_{\text{RL}+L}(\pi) = L_{\text{RL}}(\pi) + \mathbb{E}_{\tau^i \sim \Pi} \sum_t \lambda_L L_L(\pi, \tau_t^i), \quad (3)$$

where λ_L is a hyperparameter controlling the strength of the listening loss and is only nonzero on timesteps when the crewmates are asked to predict the identity of the imposter. This enables the model to optimize actions while improving its ability to identify imposters.

We can also define a purely *listening* policy, π_L that incorporates the listening loss without an RL component, therefore optimizing

$$L_{L \text{ only}}(\pi) = \mathbb{E}_{\tau^i \sim \Pi_{\text{rand}}} \sum_t \lambda_L L_L(\pi, \tau_t^i), \quad (4)$$

where Π_{rand} is a joint policy that uses π_{RWKV} for discussions and chooses gameplay actions uniformly at random.

Speaking: Reinforced Discussion Learning. So far, we have developed a policy that can learn to take effective actions in the environment with RL, and can update beliefs based on discussion messages. Now suppose that an agent is partnered with expert crewmates who already know how to parse messages from other players. How can this agent learn to construct helpful messages when it is their turn to speak?

Although our use of a supervised imposter prediction loss allows agents to learn how to interpret messages from other agents in the previous subsection, we cannot directly apply the same idea to

learning how to speak as there is no ground truth notion of effective messages. We instead improve the agents’ discussion abilities using reinforcement learning. Specifically, we grant rewards to the speaking agent based on the change in living crewmates’ beliefs on the true imposter after each message. Formally, let B_t be the sum of all living crewmates’ beliefs,

$$B_t = \sum_{k \in C_t} \pi^k(q|\tau_t^k), \quad (5)$$

where the q represents voting out the correct imposter, and C_t is the set of all living crewmates at time t . If t' is the previous belief-querying timestep, then the reward for crewmate i , who just finished speaking, is r_t^s :

$$r_t^s = B_t - B_{t'}. \quad (6)$$

Intuitively, this reward models the causal effect of each message on the task of predicting the correct imposter. The most effective message that a crewmate could send would convince other crewmates to vote out the true imposter.

Using speaking and listening, we can train an agent π_{RL+S+L} that minimizes the following loss:

$$L_{RL+L+S}(\pi) = L_{RL+L}(\pi) - \mathbb{E}_{\tau^i \sim \Pi} \sum_t [\lambda_S y^t r_t^s]. \quad (7)$$

5.3 Training for Dynamic Settings

As a team zero-sum game, we want our trained crewmates to work well against a wide range of imposters. To do so, we employ an iterated self-play algorithm, where crewmates and imposters train against earlier iterations of their adversary’s policy. We train imposters to learn to mislead crewmates into voting out other agents, so we keep the RL loss and invert the speaking loss, minimizing the following:

$$L_{\text{imp}}(\pi) = L_{\text{RL}}(\pi) + \mathbb{E}_{\tau^i \sim \Pi} \sum_t [\lambda_S y^t r_t^s]. \quad (8)$$

As the inner optimization loop, we use independent PPO [16, 32] with shared networks for policy and value functions and the Schedule Free AdamW optimizer [6].

We also want our crewmates to be robust to different partners who also act reasonably. Therefore, we always set one crewmate to be frozen to the listening policy π_L when forming the joint policy Π , following the N-Agent Ad hoc teamwork setting [37] instead of assuming a homogeneous population. This change also ensures that crewmates cannot simply determine the identity of the imposter by forming an arbitrary convention and voting out any agent who violates that convention.

Finally, we want our agents to be robust to different environment configurations. We randomize multiple environment parameters while training: choosing between three different layouts of the environment (1×3 , 2×2 , and 2×3 grids), and randomizing the number of tasks assigned to each crewmate to either 3, 4, or 5. We only train on configurations where there are 4 crewmates and 1 imposter, but we report generalization results when playing with different numbers of crewmates.

To stabilize training, we also include the following world modeling loss to each model’s loss function:

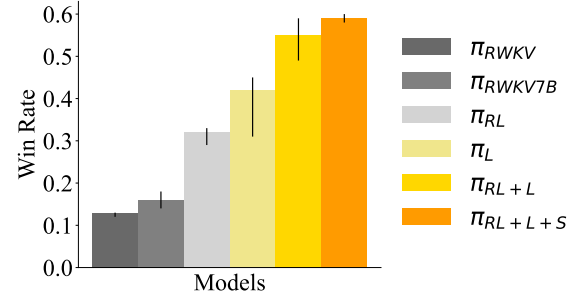


Figure 3: Win rates for crewmates trained with different algorithms over the “base” environment: 2×2 grid of rooms, 4 tasks per crewmate, and 5 players. Error bars represent the maximum and minimum expected win rates across the three independently trained runs with different seeds.

$$L_{WM}(\pi) = - \mathbb{E}_{\tau^i \sim \Pi} \sum_t \lambda_{WM} \log \pi(o_{t+1}^i | \tau_t^i, a_t^i), \quad (9)$$

where λ_{WM} is the relative strength of the world modeling loss. Although this loss does not directly contribute to improving task success, it subtly helps improve the model’s performance. In particular, as a recurrent model, RWKV benefits from this world modeling loss as it ensures that features are remembered throughout training. Furthermore, the world modeling loss prevents the model from placing too much weight on action tokens, which would cause models to output action tokens even during regular discussion sections.

6 RESULTS

In this section, we analyze the quality of our trained crewmates. We inspect the importance of different design decisions regarding the training of crewmates by ablating over the components of the loss function in Eq. (7): RL, listening, and speaking. We determine the importance of discussions in the game by measuring the equilibrium win rates of crewmates against imposters and analyze emergent behaviors in discussions. Finally, we highlight some common failure modes we observed during training and how they are mitigated in our final training algorithms.

6.1 Cooperative Training

For this set of experiments, we analyze the performance of the crewmates from the first iteration of the self-play algorithm. We conduct all experiments using the 1.5B RWKV model, because we find diminishing returns at higher parameter counts from our experiments on base models (see Appendix B for more details). We report the win rates of each policy in the base environment when keeping the imposter and one crewmate fixed to π_L in Fig. 3.

Model Evaluations. The simplest baselines are direct evaluations of the base model. We find that the 1.5B RWKV model struggles to win in the game, with the larger 7B parameter model performing slightly better as both win less than 20% of the time in the base environment.

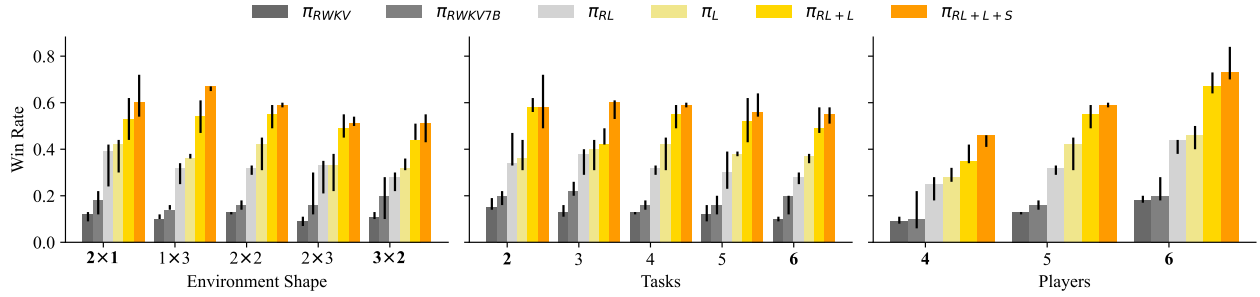


Figure 4: Win rates for crewmates trained with different algorithms over different configurations of the environment, modifying the environment shape, tasks, and number of players.

Just training with RL significantly boosts the performance relative to the base models, even significantly outperforming the 7B parameter model. However, we still find that RL without the additional listening loss struggles to reason about the identity of imposters. Even when warm-starting the RL policy from π_L , we find that it quickly loses the ability to identify imposters, instead voting for any agent with equal probability. When we instead only trained with listening – using the loss L_L only – the model, π_L , does not know which actions are effective or how to discuss details about the environment, but it is an effective baseline due to the fact that predicting the identity of the imposter is valuable in AMONG US.

When combining RL and the listening loss, we find success rates again increase dramatically, with further improvements when adding our denser *speaking* rewards, as agents can now differentiate between helpful and unhelpful messages when training. We ultimately find that our full model achieves twice the win rate of the RL-only baseline in the base environment. Note that the difference in scores when adding the additional speaking term is relatively small. Even without the explicit speaking reward, the language model produces coherent messages, often sharing their current suspicions during discussion rounds, thus benefiting from discussion even without additional rewards. This is an interesting emergent behavior as it shows that speaking is indirectly improved by training the model to listen better.

Robustness to Environment Variation. We present the win rate of crewmates against imposters across different environment configurations in Fig. 4, and see that the trends between models observed in the base environment generally persist across configurations. We find that the shape of the environment has little effect on the win rates of crewmates, with smaller environments generally being easier since it is harder for imposters to kill crewmates without witnesses. We see a general decline in performance across all models when increasing the number of tasks, because this makes it harder to win the game by completing tasks instead of voting out the imposter. Finally, we see a significant increase in win rates as the number of crewmates increase, which we expect since the crewmates can still recover from incorrectly voting out a crewmate.

We do not observe significant deviations from the expected trend lines in settings that were out of the training distribution, demonstrating how the language models can extrapolate their behaviors to unseen deviations in the configuration.

Message Evaluations. We find a major difference between the message patterns of the base RWKV model and those from π_{RL+L+S} . Most messages from the base RWKV model are often unfocused, hallucinating a wider context to the game and role-playing a crewmate. Meanwhile, crewmates using π_{RL+L+S} often directly accuse the imposter or otherwise name the imposter in their messages. In general, we find that naming an agent makes it more likely for other agents to vote against them. Furthermore, crewmates share messages that resemble environment observations that helped them judge the identity of the imposter. For instance, a crewmate may say “Player Green is leaving Room (0,1)” when the body is in Room (0,1) to indicate that Player Green was running away from the dead body, which is often correlated with being the imposter. However, the crewmates sometimes tell lies in their messages – just like humans often do when playing AMONG US. In particular, they often simply make up evidence and state whatever is most convincing to other agents to get enough votes to eject the correct imposter. Representative behavior samples are provided in Appendix C.

6.2 Robustness to Imposters

When training against frozen imposters, crewmates could come up with simple strategies that result in high win rates but are easy to overcome with a more intelligent imposter. We therefore run multiple iterations of self-play to investigate whether crewmates can use discussions even against imposters that can evolve to their policies, which we illustrate in Fig. 5. Note that in exploitability curves, we would like to see convergence upon a narrow interval for both the upper and lower bounds; a weak strategy can be easily exploited at each iteration and would therefore have both lines stay far apart and converge slowly.

We find that the crewmates’ strategies are robust to imposters trained in an adversarial fashion. In particular, we see that crewmate scores converge after only a few iterations; depending on the seed, the win rate converges to between 0.51 and 0.56 on the base environment. In fact, even the crewmates that only trained on the base model imposter are relatively strong, as can be seen by the large jump in the lower bound between iterations 0 and 1. This result implies that policies discovered by π_{RL+L+S} are very robust since even facing adversarially trained imposters does not cause a significant performance drop.

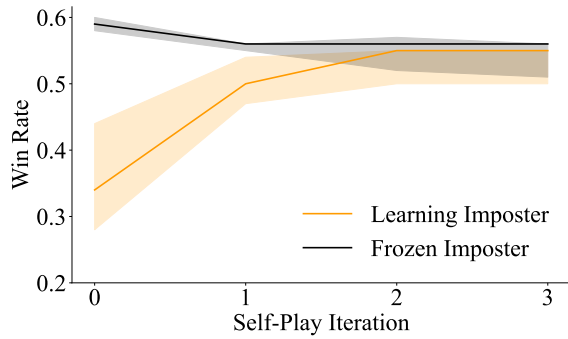


Figure 5: Exploitability curves for policies over self-play iterations, evaluated on the base environment. The orange line indicates the expected win rate against an adversarially trained imposter. The black line indicates the expected win rate of crewmates who are specifically optimized against this iteration’s imposters. Note that iteration 0 refers to the base models, while iteration 1 refers to the crewmate policy from the Cooperative Training section. Shaded regions represent the maximum and minimum win rates across the three independently trained runs with different seeds.

Qualitatively, we observe that imposters attempt to shift blame to other players by counter-accusing another crewmate. In particular, they mimic the discussion patterns of crewmates, and the crewmates sometimes fall for this deception. Crewmates who have not witnessed the murder tend to support claims made by other players, causing them to sometimes help the imposter. Interestingly, we still see similar behavior to a smaller level in the base model imposters when playing against strong crewmates. This emergent behavior can likely be attributed to the in-context learning capabilities of language models, which would allow the imposter to mimic the speech of crewmates who spoke beforehand.

6.3 Failure Modes

Throughout our experimentation, we encountered various failure modes that we tackled in our final training algorithms. Specifically, discussions tended to leave natural language and generally degenerate without careful consideration from the training algorithm.

First, we observed that the soft KL constraint commonly used in RLHF [28] required careful tuning to keep language generations in English. When this constraint is weighted too low, all of our RL-trained models diverge from natural language after only a few iterations, causing it to output random tokens during discussions and stop improving in performance.

We also observed that allowing all crewmates to be trained simultaneously would lead to degenerate solutions. Sometimes the models learn a social convention where they simply do not speak during the discussion phase. Specifically, models would output new-lines when it is their turn to speak instead of actually speaking. In this case, only the imposter would speak and all the agents would just vote the speaker out. The models would also learn to just wait in the starting room instead of moving around, allowing them to witness the murder or vote out the person who moves out of the

room. These strategies are degenerate solutions since they would not work if the imposter was aware of their strategy or if not all the crewmates shared the same strategy, but these strategies would lead to nearly perfect win rates during the first iteration of self-play. The fix to this issue was to “freeze” one crewmate to not learn and therefore not follow changes in strategies.

The final failure mode we observed was using action tokens in discussions instead of natural language. Specifically, the RL-trained models would learn to take actions by explicitly choosing action tokens, but this gives action tokens a higher probability to be chosen overall, even during discussion phases. We observed that the best way to counteract this effect was to introduce the world modeling loss from Eq. (9). This loss ensured that the model preserved its language modeling abilities and had the side effect of helping the models match the patterns it experienced in observations within its own discussions, which would help independent agents understand the intentions of our models.

7 DISCUSSION

Summary. We introduce a technique to self-improve the discussion ability of an LLM in social deduction games and show how it enables agents to communicate effectively in the game of AMONG Us. We demonstrate that, despite having weak base models, our agents learn to speak effectively and extract information from discussion messages. We also find that our agents are robust to adversarially trained imposters, who, despite attempting to sabotage the discussion, are unable to break the crewmates’ coordination during discussions. Our technique ultimately shows that self-improving discussions in multi-agent settings does not require task-specific human data, unlocking the possibility for multi-agent communication with language models in novel tasks.

Limitations and Future Work. A key limitation of our approach is that our scene prediction technique is task-dependent. In AMONG Us, there is a natural connection between the discussion and trying to predict the identity of the imposter, and a similar structure applies to a wide range of social deduction games and real-world settings. An interesting future direction would be to allow agents to identify which aspects of the scene are relevant to a specific task instead of manually specifying it. Please refer to Appendix D for more analysis on the broader impacts of our work.

We also note that crewmates are not always truthful in their discussions, opting instead to make the most convincing statements. We consider this behavior to be potentially dangerous outside of our sandboxed setting of Among Us, so we believe that optimizing for truthfulness is an important future direction.

ACKNOWLEDGMENTS

This research was supported in part by the Other Transaction award HR00112490375 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, the Cooperative AI foundation, ONR project N00014-21-1-2298, NSF Awards #2006388, #1941722, #2125511, AFOSR YIP, and the Stanford Center for Human-Centered AI (HAI). We thank the Stanford Madrona team for giving advice on the systems-level details of our implementation, and Hengyuan Hu for his insightful feedback when reviewing this paper.

REFERENCES

- [1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do As I Can and Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691 [cs.RO]
- [2] Mark Braverman, Omid Etesami, and Elchanan Mossel. 2008. Mafia: A Theoretical Study of Players and Coalitions in a Partial Information Environment. *The Annals of Applied Probability* 18, 3 (2008), 825–846. <http://www.jstor.org/stable/25442651>
- [3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Scott Lundberg, Harsha Nari, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]
- [4] Luca Carminati, Brian Hu Zhang, Gabriele Farina, Nicola Gatti, and Tuomas Sandholm. 2023. Hidden-Role Games: Equilibrium Concepts and Computation. arXiv:2308.16017 [cs.GT]
- [5] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2019. *On the utility of learning about humans for human-AI coordination*. Curran Associates Inc., Red Hook, NY, USA.
- [6] Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. 2024. The Road Less Scheduled. In *Thirty-eighth Conference on Neural Information Processing Systems*.
- [7] Yujian Dong, Tianyu Wu, and Chaoyang Song. 2024. Optimizing Robotic Manipulation with Decision-RWKV: A Recurrent Sequence Modeling Approach for Lifelong Learning. arXiv:2407.16306 [cs.RO] <https://arxiv.org/abs/2407.16306>
- [8] FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074. <https://doi.org/10.1126/science.ade9097> arXiv:<https://www.science.org/doi/pdf/10.1126/science.ade9097>
- [9] Michael C. Frank and Noah D. Goodman. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology* 75 (2014), 80–96. <https://doi.org/10.1016/j.cogpsych.2014.08.002>
- [10] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Li Fei-Fei, Jianfeng Gao, and Hoi Vo. 2024. MindAgent: Emergent Gaming Interaction. 3154–3183. <https://doi.org/10.18653/v1/2024.findings-naacl.200>
- [11] Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: learning to communicate with sequences of symbols. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) (NIPS’17). Curran Associates Inc., Red Hook, NY, USA, 2146–2156.
- [12] Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual Adaptation for Efficient Machine Communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Raquel Fernández and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 408–419. <https://doi.org/10.18653/v1/2020.conll-1.33>
- [13] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. “Other-Play” for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning (ICML’20)*. JMLR.org, Article 409, 12 pages.
- [14] Hengyuan Hu and Dorsa Sadigh. 2023. Language Instructed Reinforcement Learning for Human-AI Coordination. In *40th International Conference on Machine Learning (ICML)*.
- [15] Jerry Huang. 2024. How Well Can a Long Sequence Model Long Sequences? Comparing Architectural Inductive Biases on Long-Context Abilities. arXiv:2407.08112 [cs.LG] <https://arxiv.org/abs/2407.08112>
- [16] Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. 2022. CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms. *Journal of Machine Learning Research* 23, 274 (2022), 1–18. <http://jmlr.org/papers/v23/21-1342.html>
- [17] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 9118–9147. <https://proceedings.mlr.press/v162/huang22a.html>
- [18] Innersloth. 2024. Among Us. <https://www.innersloth.com/games/among-us/>. [Online; accessed 25-February-2024].
- [19] Kavya Koppurapu, Edgar A. Duéñez-Guzmán, Jayd Matyas, Alexander Sasha Vezhnevets, John P. Agapiou, Kevin R. McKee, Richard Everett, Janusz Marecki, Joel Z. Leibo, and Thore Graepel. 2022. Hidden Agenda: a Social Deduction Game with Diverse Learned Equilibria. arXiv:2201.01816 [cs.AI]
- [20] Minae Kwon, Hengyuan Hu, Vivek Myers, Siddharth Karamcheti, Anca Dragan, and Dorsa Sadigh. 2024. Toward Grounded Social Reasoning. In *International Conference on Robotics and Automation (ICRA)*.
- [21] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward Design with Language Models. In *International Conference on Learning Representations (ICLR)*.
- [22] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Reh, and Diyi Yang. 2023. Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6570–6588. <https://doi.org/10.18653/v1/2023.findings-acl.411>
- [23] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Hk8N3ScIq>
- [24] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023. Text2Motion: from natural language instructions to feasible plans. *Autonomous Robots* (14 Nov 2023). <https://doi.org/10.1007/s10514-023-10131-7>
- [25] Qinghua Liu, Csaba Szepesvari, and Chi Jin. 2022. Sample-Efficient Reinforcement Learning of Partially Observable Markov Games. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=HnIQR5Y7vPI>
- [26] William P. McCarthy, Robert D. Hawkins, Haoliang Wang, Cameron Holdaway, and Judith E. Fan. 2021. Learning to communicate about shared procedural abstractions. arXiv:2107.00077 [cs.CL]
- [27] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large Language Models as General Pattern Machines. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*.
- [28] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]
- [29] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)* (San Francisco, CA, USA) (UIST ’23). Association for Computing Machinery, New York, NY, USA.
- [30] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocan, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the Transformer Era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14048–14077. <https://doi.org/10.18653/v1/2023.findings-emnlp.936>
- [31] Bidipta Sarkar, Andy Shih, and Dorsa Sadigh. 2024. Diverse conventions for human-AI collaboration. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS ’23). Curran Associates Inc., Red Hook, NY, USA, Article 1003, 25 pages.
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG]
- [33] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. arXiv:2305.14763 [cs.CL]
- [34] Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. GPT-4 Doesn’t Know It’s Wrong: An Analysis of Iterative Prompting for Reasoning Problems. arXiv:2310.12397 [cs.AI]
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] <https://arxiv.org/abs/1706.03762>
- [36] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky,

- James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575 (2019), 350 – 354. <https://api.semanticscholar.org/CorpusID:204972004>
- [37] Caroline Wang, Arrasy Rahman, Ishan Durugkar, Elad Liebman, and Peter Stone. 2024. N-Agent Ad Hoc Teamwork. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [38] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv preprint arXiv: Arxiv-2305.16291* (2023).
- [39] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-Rewarding Language Models. arXiv:2401.10020 [cs.CL]