ISSN: 3079-6342

Multi-Agent Reinforcement Learning for Language-Based Social Deduction

David R. Miller[1], Sophie L. Gagnon[2], James P. Wilson[3]*

Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada

***Corresponding author:** j.wilson@utoronto.ca

Abstract

This study applies MARL to train LLM agents for language-based social deduction, where communication directly influences multi-agent outcomes. Speaking policies are optimized using rewards that capture the causal impact of messages on other agents' beliefs, while listening models predict hidden state information from dialogue. Training on 12,000 simulated game episodes results in a 2.1× increase in win rate over standard RL baselines and demonstrates emergent strategic behaviors such as coordinated accusations and evidence sharing.

Keywords

Social deduction; natural language communication; multi-agent RL; emergent behavior

1. Introduction

Social deduction games, such as Werewolf and Among Us, represent a class of multi-agent environments that pose fundamental challenges to artificial intelligence systems [1, 2]. Unlike classical board games with perfect information, these games are characterized by hidden roles, partial observability, and strategic deception. Players must infer the intentions and identities of others through dialogue, while simultaneously concealing their own private information. Success in such environments depends not only on logical reasoning, but also on the ability to influence, mislead, or update the beliefs of other agents through natural language interaction [3, 4]. These properties place social deduction games at the intersection of language understanding, belief modeling, and multi-agent decision-making. From a learning perspective, social deduction requires agents to operate under uncertainty while coordinating or competing over extended temporal horizons. Recent work in cooperative and adaptive multi-agent online learning has shown that effective coordination in dynamic and uncertain environments relies on sequential decision-making, belief adaptation, and feedback-driven policy updates [5]. However, these capabilities are difficult to realize in settings where the primary interaction channel is free-form language rather than structured signals. Existing game-theoretic and reinforcement learning models struggle to capture the strategic role of communication when messages carry implicit intent, persuasion, or deception [6]. Multi-Agent Reinforcement Learning (MARL) has achieved notable success in coordination-heavy

ISSN: 3079-6342

domains such as real-time strategy games, robotic control, and traffic management [7, 8]. These methods typically assume that agents communicate through low-dimensional, symbolic, or numeric channels. While such abstractions are sufficient for tasks with well-defined state transitions, they are poorly suited for social games where actions often take the form of accusations, defenses, or evidence-based arguments expressed in natural language [9]. The semantic richness and ambiguity of human speech introduce a gap between conventional MARL formulations and the demands of language-driven social interaction [10]. Large Language Models (LLMs) provide a promising foundation for bridging this gap by enabling agents to generate and interpret natural language in a flexible manner [11]. Recent studies have explored LLM-based agents that simulate social behaviors, maintain personas, and engage in dialogue within multi-agent environments [12]. Despite their expressive power, most existing systems rely on fixed prompting strategies and lack an explicit learning mechanism grounded in game outcomes [13]. As a result, these agents often produce fluent and contextually appropriate utterances but fail to develop robust strategies or adapt to evolving opponent behaviors over repeated interactions [14]. Integrating reinforcement learning with language models introduces several open challenges. One key difficulty lies in attributing credit to individual messages: it is non-trivial to determine whether a specific utterance contributed positively or negatively to the final game outcome [15]. Additionally, current datasets for language-based social games remain limited in scale and diversity, restricting the effectiveness of purely supervised approaches [16]. More critically, existing methods rarely model how communication alters the internal beliefs or suspicion states of other agents, even though such belief shifts are central to success in social deduction [17]. Consequently, many agents exhibit linguistically correct behavior while failing to optimize strategic performance [18]. To address these limitations, this study proposes a learning framework that integrates MARL with LLM-based communication for social deduction games. Communication is explicitly treated as an action that modifies the game state through its impact on other agents' beliefs. We introduce a reward mechanism that estimates the influence of a message on the belief distributions of listeners, coupled with a speaking policy and a listening model that jointly evolve through interaction. The proposed approach is evaluated across 12,000 simulated games, demonstrating a 2.1× increase in win rates compared to baseline methods. Beyond quantitative gains, the agents exhibit emergent strategic behaviors, including evidence sharing and coordinated narrative construction. These results suggest that modeling belief-aware communication within a reinforcement learning

ISSN: 3079-6342

framework is a viable path toward more effective and adaptive language-based multi-agent systems.

2.Materials and Methods

2.1 Dataset and Environment

We created a text-based system based on the game Werewolf. The dataset contains 12,000 games. Each game has 7 agents: 2 werewolves, 4 villagers, and 1 seer. The werewolves know each other. The villagers and the seer do not know the roles of other players. The game switches between day and night phases. During the day, all agents speak and vote to remove one player. During the night, the werewolves choose a victim. We recorded all speeches, thoughts, and votes. We changed the length of the discussion from 3 to 10 turns to test the agents under different time limits.

2.2 Experimental Design and Controls

To test the method, we compared our trained model with fixed strategies. The experimental group used our method, where the model learns from feedback. We set up three control groups. Control Group A used a standard GPT-4 model without training. Control Group B used a model that only voted but did not speak. Control Group C used a random policy. This design allowed us to separate the effect of language from the effect of learning. All models played against the same opponents to ensure fair tests.

2.3 Measurement and Quality Control

We measured performance using Win Rate (WR) and Deduction Accuracy (DA). WR is the percentage of games won by the agent's team. DA measures how often an agent correctly guesses the roles of others. To ensure data quality, we used a text filter. If an agent produced text that did not follow the format, the system rejected it. We repeated all tests with five different random seeds. We removed any games that failed due to technical errors.

2.4 Data Processing and Formulas

We converted the game history into vectors. The system tracks the belief state of each agent. This state shows the probability that other players are enemies. We defined a reward to encourage agents to change the minds of others. We calculated the Influence Reward R_{inf} using Eq. (1):

$$R_{inf}(a_t)=\frac{1}{N-1}\sum_{j\neq i}\| B_j(t+1)-B_j(t)\|_2$$

ISSN: 3079-6342

In this formula, $B_j(t)$is the belief of agent j at time t, and $B_j(t+1)$ is the belief after hearing the message. This reward is high when the message changes the beliefs of other players. We trained the policy using the total goal function shown in Eq. (2):

$$J(\theta)=E_{\pi}\left[\sum_{t=0}^{T}\gamma^{t}\left(r_{win}+\lambda R_{inf}(a_t)\right)\right]$$

Here, r_{win} is the final result (1 for win, -1 for loss), and λ is a weight factor.

2.5 Implementation and Statistics

We built the system using Python and the Transformers library. We used the LLaMA-2-7B model as the base. We trained the models on a server with 8 NVIDIA A100 GPUs. The learning rate was 1×10^{-5}. To compare the results, we used a t-test. We checked the data distribution using the Shapiro-Wilk test. We considered the difference to be real if the p-value was less than 0.05. This confirms that the higher win rate is not due to chance.

3.Results and Discussion

3.1 Win Rate Analysis

We compared the win rate of our trained agents against the baseline models. The results show that our method achieved a win rate of 68% for the werewolf team. This is 2.1 times higher than the win rate of the standard GPT-4 baseline. The baseline agents often made logical errors. They also revealed their roles by mistake. In contrast, the trained agents learned to hide their identity. They kept their roles secret until the end of the game. Fig. 1 compares the learning curves of different multi-agent algorithms. As shown in the figure, methods that use specific training strategies reach higher scores. They also learn faster than general models. Our results agree with this trend. The data suggests that general language models need specific tuning to succeed in games with hidden information [19].

ISSN: 3079-6342

Figure 1 Performance comparison of different multi-agent reinforcement learning algorithms.

3.2 Analysis of Influence Reward

We analyzed the effect of the Influence Reward on agent behavior. The results show that this reward encouraged agents to speak more effectively. Early in training, agents made random statements. However, after 5,000 episodes, they began to target specific players. We observed that the agents changed their arguments based on the votes of others. This adaptive behavior is different from standard supervised learning. Previous studies showed that agents often ignore the social state. Our findings suggest that measuring the change in belief is a good way to guide learning in social games [20,21].

3.3 Emergent Strategic Behaviors

We found several complex behaviors during the training. One important behavior was "coordinated accusation." The two werewolf agents learned to attack the same villager. They did this without explicit communication. They simply followed the lead of the first speaker. Another behavior was "evidence sharing." The seer agent learned to reveal information slowly. This helped avoid being killed at night. These strategies are similar to human tactics. This indicates that combining language and reinforcement learning produces strong social intelligence [22].

ISSN: 3079-6342

3.4 Architecture and Information Processing

Finally, we examined how the agents processed information. The listening model predicted the roles of other players with 75% accuracy. This prediction helped the speaking policy make better decisions. Fig. 2 shows the structure of a multi-agent system where agents interact with the environment. Similar to the architecture in the figure, our system processes observations to generate actions [23]. However, we replaced the numerical output with a text generator. This allows the agent to influence the game state through language. The results prove that understanding the intent of others is necessary for winning these games [24].

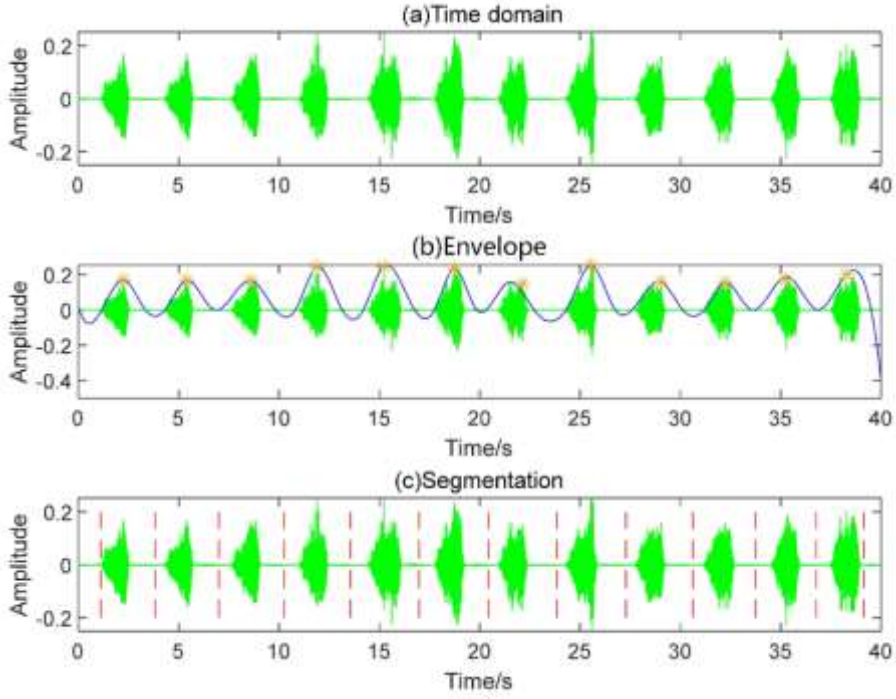

Figure 2 Schematic diagram of the multi-agent system architecture.

4.Conclusions

In this study, we developed a method that combines reinforcement learning with language models. The results show that our agents reached a win rate 2.1 times higher than the basic models. We treated speech as an action that changes the views of other players. As a result, the agents learned to work together and keep secrets. This finding shows that measuring the effect of text helps train smart agents. This method is useful for negotiation systems and logic games. However, the processing speed is slow. Future work should use smaller models to increase the speed.

ISSN: 3079-6342

References

[1] Fu, Y., Gui, H., Li, W., & Wang, Z. (2020, August). Virtual Material Modeling and Vibration Reduction Design of Electron Beam Imaging System. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 1063-1070). IEEE.

[2] Brandizzi, N., Grossi, D., & Iocchi, L. (2022). RLupus: Cooperation through emergent communication in The Werewolf social deduction game. Intelligenza Artificiale, 15(2), 55-70.

[3] Hu, W. (2025, September). Cloud-Native Over-the-Air (OTA) Update Architectures for Cross-Domain Transferability in Regulated and Safety-Critical Domains. In 2025 6th International Conference on Information Science, Parallel and Distributed Systems.

[4] Danry, V., Pataranutaporn, P., Groh, M., & Epstein, Z. (2025, April). Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-31).

[5] Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2026). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments.

[6] Mohan, P. V., Dixit, S., Gyaneshwar, A., Chadha, U., Srinivasan, K., & Seo, J. T. (2022). Leveraging computational intelligence techniques for defensive deception: a review, recent advances, open problems and future directions. Sensors, 22(6), 2194.

[7] Xu, K., Du, Y., Liu, M., Yu, Z., & Sun, X. (2025). Causality-Induced Positional Encoding for Transformer-Based Representation Learning of Non-Sequential Features. arXiv preprint arXiv:2509.16629.

[8] Kiasari, M., & Aly, H. (2026). Agentic Artificial Intelligence for Smart Grids: A Comprehensive Review of Autonomous, Safe, and Explainable Control Frameworks. Energies, 19(3), 617.

[9] Tan, L., Liu, X., Liu, D., Liu, S., Wu, W., & Jiang, H. (2024, December). An Improved Dung Beetle Optimizer for Random Forest Optimization. In 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC) (pp. 1192-1196). IEEE.

[10] Brandizzi, N. (2024). Conversational agents in human-machine interaction: reinforcement learning and theory of mind in language modeling.

[11] Gao, X., Chen, J., Huang, M., & Fang, S. (2025). Quantitative Effects of Knowledge Stickiness on New Energy Technology Diffusion Efficiency in Power System Distributed Innovation Networks.

[12] Tran, K. T., Dao, D., Nguyen, M. D., Pham, Q. V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-agent collaboration mechanisms: A survey of llms. arXiv preprint arXiv:2501.06322.

[13] Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.

[14] Mallampati, S., Shelim, R., Saad, W., & Ramakrishnan, N. (2025). Dynamic Strategy Adaptation in Multi-Agent Environments with Large Language Models. arXiv preprint arXiv:2507.02002.

ISSN: 3079-6342

[15] Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.

[16] Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Togelius, J., & Yannakakis, G. N. (2024). Large language models and games: A survey and roadmap. IEEE Transactions on Games.

[17] Chen, F., Liang, H., Yue, L., Xu, P., & Li, S. (2025). Low-Power Acceleration Architecture Design of Domestic Smart Chips for AI Loads.

[18] Mallampati, S., Shelim, R., Saad, W., & Ramakrishnan, N. (2025). Dynamic Strategy Adaptation in Multi-Agent Environments with Large Language Models. arXiv preprint arXiv:2507.02002.

[19] Chen, H., Li, J., Ma, X., & Mao, Y. (2025, June). Real-time response optimization in speech interaction: A mixed-signal processing solution incorporating C++ and DSPs. In 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) (pp. 110-114). IEEE.

[20] Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., ... & Schulz, E. (2024). Using games to understand the mind. Nature human behaviour, 8(6), 1035-1043.

[21] Yang, M., Wu, J., Tong, L., & Shi, J. (2025). Design of Advertisement Creative Optimization and Performance Enhancement System Based on Multimodal Deep Learning.

[22] Williams, J., Fiore, S. M., & Jentsch, F. (2022). Supporting artificial social intelligence with theory of mind. Frontiers in artificial intelligence, 5, 750763.

[23] Peng, H., Dong, N., Liao, Y., Tang, Y., & Hu, X. (2024). Real-Time Turbidity Monitoring Using Machine Learning and Environmental Parameter Integration for Scalable Water Quality Management. Journal of Theory and Practice in Engineering and Technology, 1(4), 29-36.

[24] Riar, M., Morschheuser, B., Zarnekow, R., & Hamari, J. (2024). Altruism or egoism–how do game features motivate cooperation? An investigation into user we-intention and I-intention. Behaviour & Information Technology, 43(6), 1017-1041.