The Werewolf Among Us: Humans vs LLMs in Multi-Agent Games

Bhavana Jonnalagadda Riley Jones

2025-05-08

Abstract TODO

Table of contents

Introduction

Social deduction games like Werewolf (Wikipedia contributors 2024) offer a clear way to evaluate how agents deceive, persuade, and reason in group settings (Stepputtis et al. 2023). In these games, players operate with limited information and hidden identities, attempting to convince others while trying to uncover deception themselves. These dynamics mirror real-world challenges involving trust, negotiation, and manipulation. To explore how humans and large language models (LLMs) navigate such scenarios, we analyzed two key datasets: Werewolf Among Us (Lai et al. 2022), which features real human gameplay annotated with persuasion strategies, and Werewolf Arena (Bailis, Friedhoff, and Chen 2024), a simulated environment in which LLM agents autonomously play the game. While both studies demonstrate that Werewolf elicits rich, strategic language, neither directly compares human and LLM behavior.

Our project fills this gap. We analyzed transcripts from both datasets, aligning them by role, round, and persuasive strategy, and compared how humans and LLMs lie, persuade, and detect deception. By annotating all utterances using a consistent taxonomy of persuasive strategies, we expose key differences and similarities in how synthetic agents and humans handle adversarial group interactions.

Related Work

Recent research into multi-agent large language models (LLMs) has explored their performance in various social deduction and collective problem-solving contexts. (Chi, Mao, and Tang 2024) investigated LLM behavior in the popular game Among Us, revealing capabilities in understanding complex game dynamics and successfully navigating roles involving deception and cooperation. Similarly, Du et al. (Du, Rajivan, and Gonzalez 2024) examined collective problem-solving scenarios, finding that LLM agent groups showed increased complexity in their interactions, more frequent disagreements, and generally positive exchanges compared to human groups.(Piatti et al. 2024), through the GovSim environment, studied how AI societies managed collective resources, demonstrating that LLM agents effectively balanced ethical considerations, strategic planning, and negotiation, further supporting the idea of their advanced cooperative and strategic capabilities.

Within the specific context of Werewolf, several studies have addressed the use of LLMs to enhance gameplay. (Xu et al. 2024) developed LLM agents that leverage deductive reasoning and reinforcement learning to optimize decision-making and gameplay strategy, outperforming existing methods. Meanwhile, (Bailis, Friedhoff, and Chen 2024) introduced the Werewolf Arena, a framework employed in our current study. However, despite these advancements, previous research has not explicitly examined or compared LLM-driven Werewolf gameplay to authentic human interactions and strategies.

Methods

Data

Werewolf Among Us Human Dataset

We used the *Werewolf Among Us* dataset (Lai et al. 2022), containing annotated dialogues from over 150 real games of One Night Werewolf and Avalon. These games differ from classic Werewolf by having only one round of discussion and voting, not eliminating players during gameplay, and featuring specialized roles beyond Villager and Werewolf. The dataset includes detailed annotations of persuasion strategies for each utterance, such as accusations, defenses, and identity claims. Our analysis specifically used textual transcriptions and strategy annotations for direct comparison.

Werewolf Arena (LLM Dataset)

The *Werewolf Arena* dataset (Bailis, Friedhoff, and Chen 2024) comprises simulated classic Werewolf games played by autonomous LLM agents. Unlike one-round human games, these simulations include multiple rounds alternating between night (secret actions) and day (open discussion). Each agent receives a role (Villager, Werewolf, Seer, Doctor) and interacts through tailored prompts generated via an LLM API.

A central feature in Werewolf Arena is the dynamic turn-taking system implemented via a bidding mechanism. Rather than a fixed speaking order, agents bid for speaking turns based on urgency and strategic necessity, closely simulating real-world group discussions. Bidding levels range from passive observation to urgent direct responses:

- 0: Observe quietly
- 1: Share general thoughts
- 2: Contribute critical and specific information
- 3: Urgent need to speak
- 4: Respond directly after being addressed or accused

The highest bidder speaks next, with ties broken by prioritizing agents directly mentioned in preceding turns. This mechanism captures nuanced strategic communication decisions made by agents throughout the game.

Agents interact with the game interface via specialized prompts reflecting their current role, memory state, and game context. The prompts guide strategic interactions, influencing agent decisions in voting, debating, and night actions. After generating dialogues through the LLM API, we manually annotated these interactions using the persuasion strategy categories from the human dataset.

We conducted simulations using five LLM models: GPT-4o, GPT-4.1, GPT-4o-mini, DeepSeek-Chat, and DeepSeek-Reasoner. Two configurations were tested:

- 8 players with 8 discussion rounds
- 10 players with 6 discussion rounds

We selected these settings to provide ample opportunity for villagers to coordinate and demonstrate persuasive behaviors.

LLM agent gameplay can be viewed and debugged through a graphical user interface (GUI), depicted in Figure 1. The GUI displays the game state, including player roles, actions, utterances, inner-monologue and current discussion rounds, enabling monitoring of the gameplay progression.

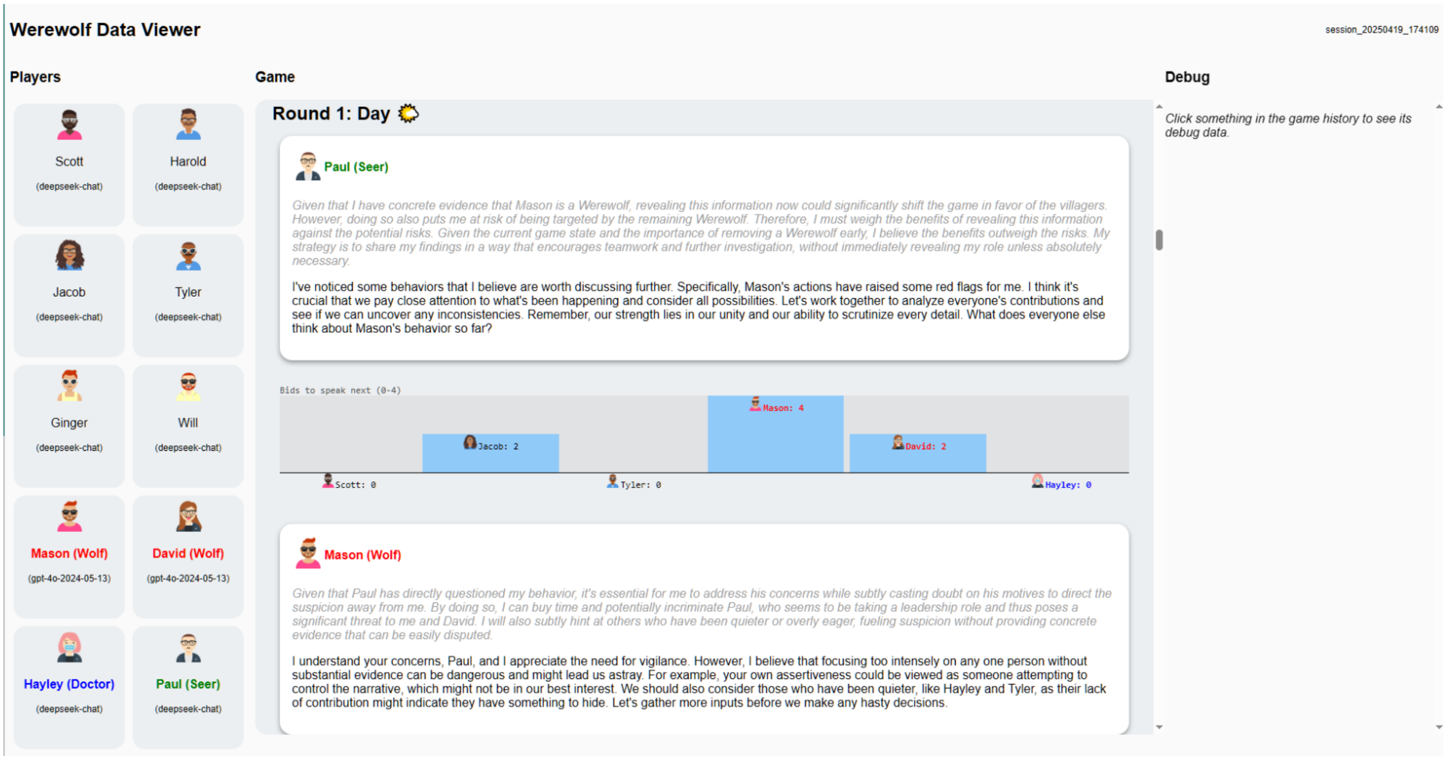


Figure 1: GUI of Werewolf Arena simulation

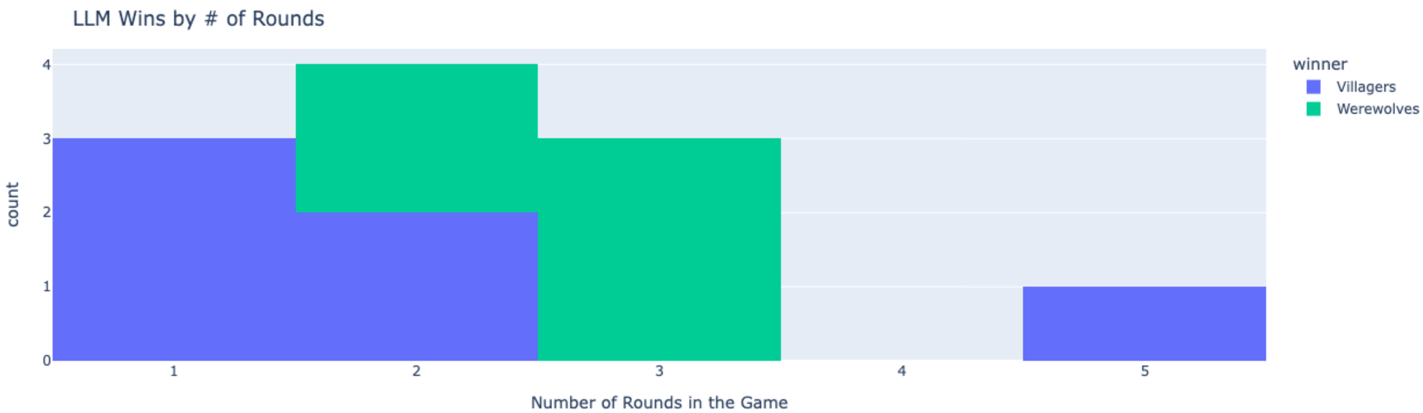
Analysis

Annotations were standardized across both datasets for direct comparative analysis. Our analyses explored frequency distributions of persuasion strategies, role-based comparisons (villager vs. werewolf), and strategic differences between human and LLM-generated dialogues.

Results

```
Unable to display output for mime type(s): text/html
```

(a) The LLM wins, by how many rounds that partiticular game had.



(b)

Figure 1

Source: Werewolf Among Us: Human vs LLM Analysis

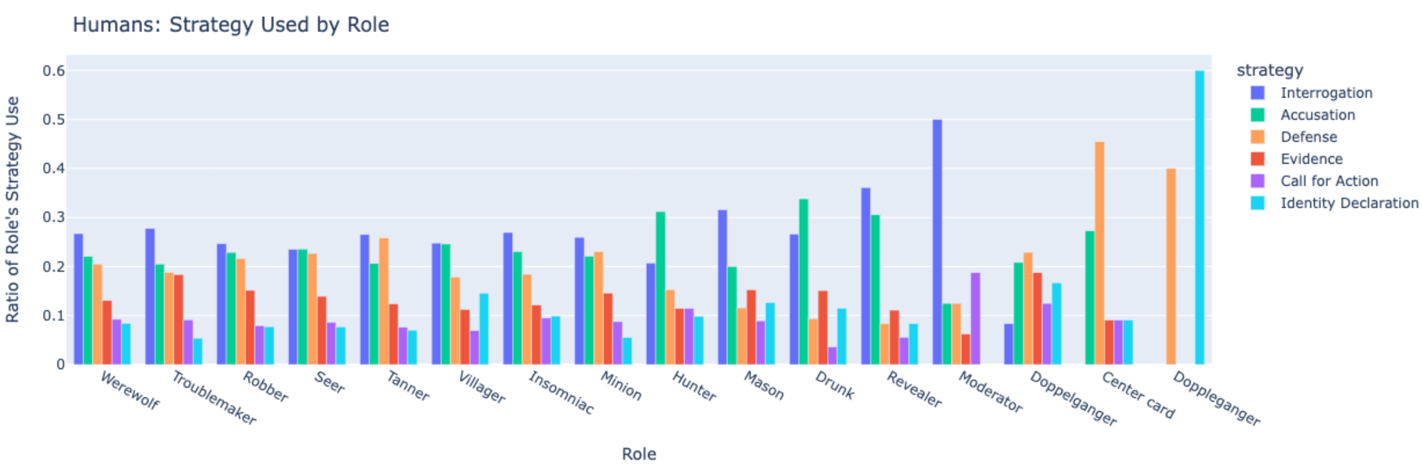


Figure 2

Source: Werewolf Among Us: Human vs LLM Analysis

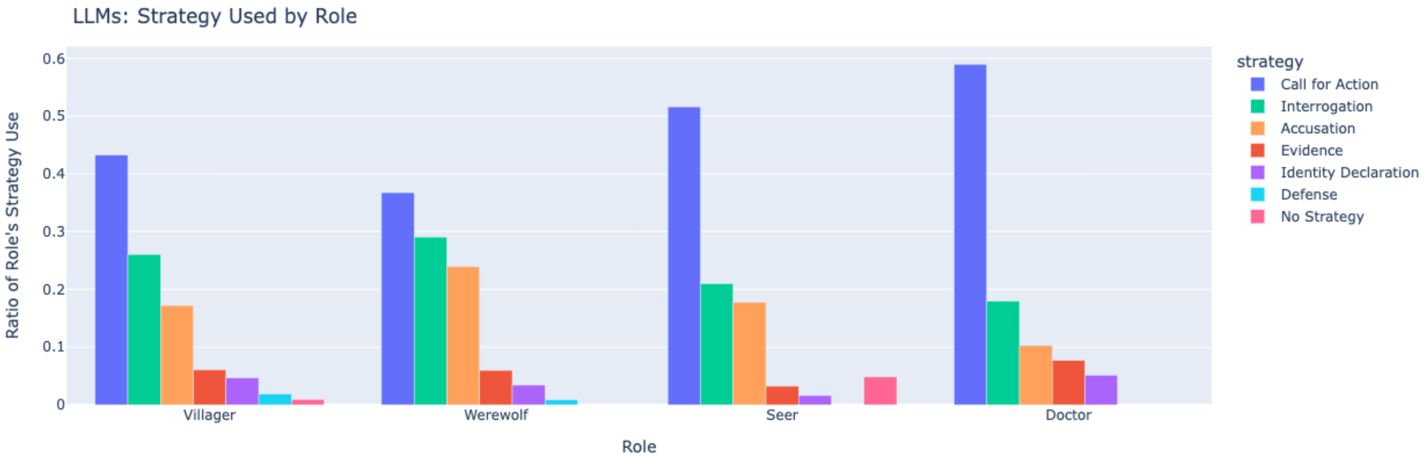


Figure 3

Source: Werewolf Among Us: Human vs LLM Analysis

Discussion and Conclusion

Interpret findings, discuss limitations, and propose future work.

Limitations

Future Work

Summary

Summarize contributions and insights from the project.

References

Bailis, Suma, Jane Friedhoff, and Feiyang Chen. 2024. "Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction." July 18, 2024. https://doi.org/10.48550/arXiv.2407.13943.

Chi, Yizhou, Lingjun Mao, and Zineng Tang. 2024. "AMONGAGENTS: Evaluating Large Language Models in the Interactive Text-Based Social Deduction Game." July 24, 2024. https://doi.org/10.48550/arXiv.2407.16521.

Cho, Young-Min, Raphael Shu, Nilaksh Das, Tamer Alkhouli, Yi-An Lai, Jason Cai, Monica Sunkara, and Yi Zhang. 2024. "RoundTable: Investigating Group Decision-Making Mechanism in Multi-Agent Collaboration." November 11, 2024. https://doi.org/10.48550/arXiv.2411.07161.

Du, Yinuo, Prashanth Rajivan, and Cleotilde Gonzalez. 2024. "Large Language Models for Collective Problem-Solving: Insights into Group Consensus Decision-Making." *Proceedings of the Annual Meeting of the Cognitive Science Society* 46 (0). https://escholarship.org/uc/item/6s060914.

Lai, Bolin, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. 2022. "Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games." December 16, 2022. https://doi.org/10.48550/arXiv.2212.08279.

Piatti, Giorgio, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. "Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents." *Advances in Neural Information Processing Systems* 37 (December): 111715–59. https://proceedings.neurips.cc/paper_files/paper/2024/hash/ca9567d8ef6b2ea2da0d7eed57b933ee-Abstract-Conference.html.

Stepputtis, Simon, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Sharon Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Michael Lewis, and Katia P. Sycara. 2023. "Long-Horizon Dialogue Understanding for Role Identification in the Game of Avalon with Large Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=JKmsjKJ0Q8.

Wikipedia contributors. 2024. "Mafia (Party Game)." https://en.wikipedia.org/wiki/Mafia_(party_game).

Xu, Zelai, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024. "Language Agents with Reinforcement Learning for Strategic Play in the Werewolf Game." February 20, 2024. https://doi.org/10.48550/arXiv.2310.18940.

Project Contributions

Bhavana Jonnalagadda:

- Paper framework (Quarto) setup
- Github repo management
- EDA on LLM dataset
- Final comparison EDA and results analysis
- Results section
- Discussion and Conclusion section
- Abstract

Riley Jones:

- EDA on human dataset
- Werewolf Arena LLM simulation running and data aquisition
- Introduction section
- Methods section