

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

Bhavana Jonnalagadda

Riley Jones

2025-05-07

Abstract TODO

Table of contents

Introduction	4
Related Work	4
Multi-Agent LLMs	4
LLMs and Werewolf	4
Methods	6
Data	6
Werewolf Among Us Human Dataset	6
Werewolf Arena (LLM Dataset)	6
Analysis	7
Results	8
Discussion and Conclusion	10
Limitations	10
Future Work	10
Summary	10
References	11
Project Contributions	12

Introduction

Social deduction games like Werewolf offer a clear way to evaluate how agents deceive, persuade, and reason in group settings ([Wikipedia contributors 2024](#)). 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

Multi-Agent LLMs

- Among us game ([Chi, Mao, and Tang 2024](#))
- Collective problem solving ([Du, Rajivan, and Gonzalez 2024](#))
 - “analyses indicate that LLM agent groups exhibit more disagreements, complex statements, and a propensity for positive statements compared to human groups”
- Govsim ([Piatti et al. 2024](#))
 - “In GOVSIM, a society of AI agents must collectively balance exploiting a common resource with sustaining it for future use. This environment enables the study of how ethical considerations, strategic planning, and negotiation skills impact cooperative outcomes.”
- All found similar themes
 - That LLMs are capable and good at understanding the rules
 - That they can cooperate and be sneaky

LLMs and Werewolf

- Examination of improving werewolf by LLMs ([Xu et al. 2024](#))
 - “our agents use an LLM to perform deductive reasoning and generate a diverse set of action candidates. Then an RL policy trained to optimize the decision-making ability chooses an action from the candidates to play in the game. Extensive experiments show that our agents overcome the intrinsic bias and outperform existing LLM-based agents in the Werewolf game.”
- Werewolf Arena ([Bailis, Friedhoff, and Chen 2024](#))
 - Used in this paper

- Explicitly discuss how none of the existing LLM+Werewolf papers examine the differences/compare from a human dataset

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 utilized 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.

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.

Agents participated in gameplay through a graphical user interface (GUI), depicted in Figure Figure 1. The GUI displays the game state, including player roles, actions, and current discussion rounds, enabling monitoring of the gameplay progression.

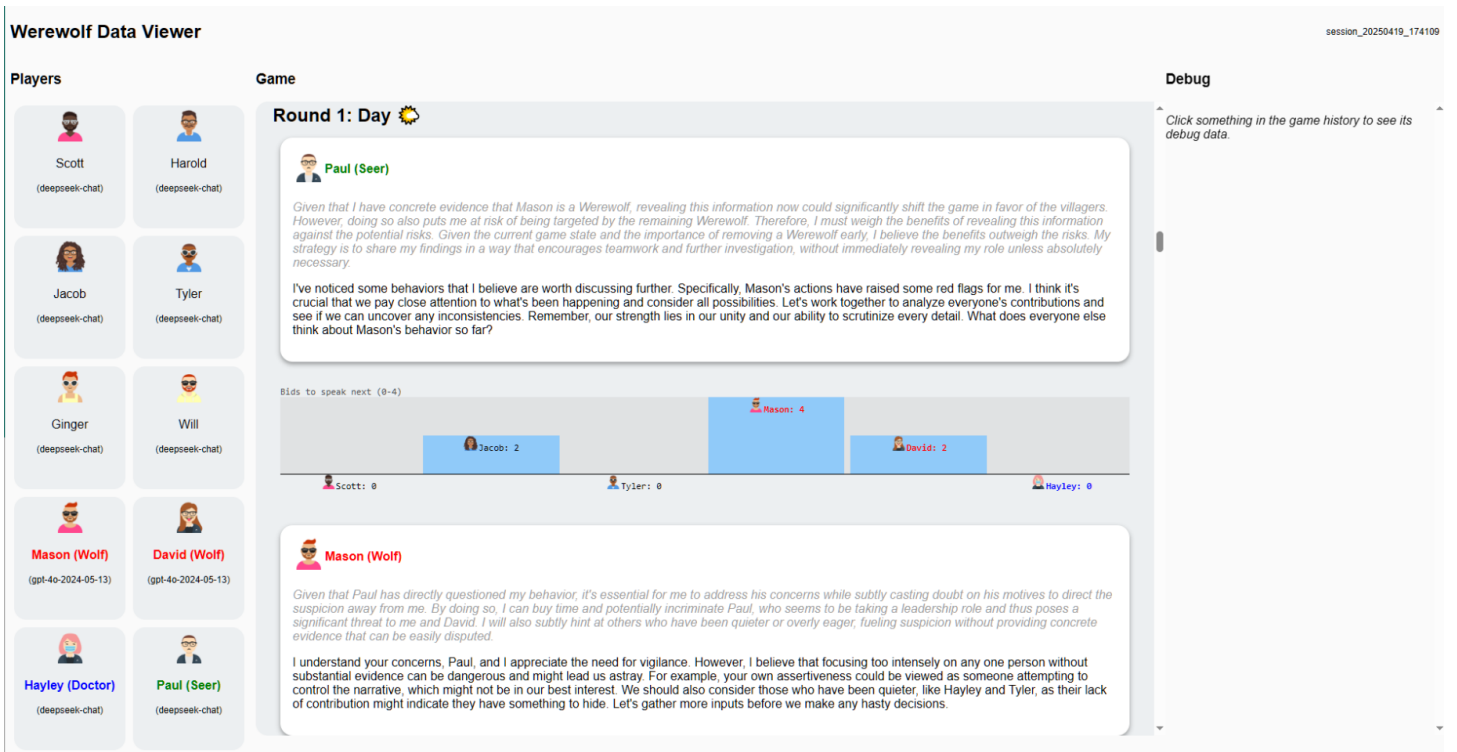


Figure 1: GUI of Werewolf Arena simulation

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 utilize 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.

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

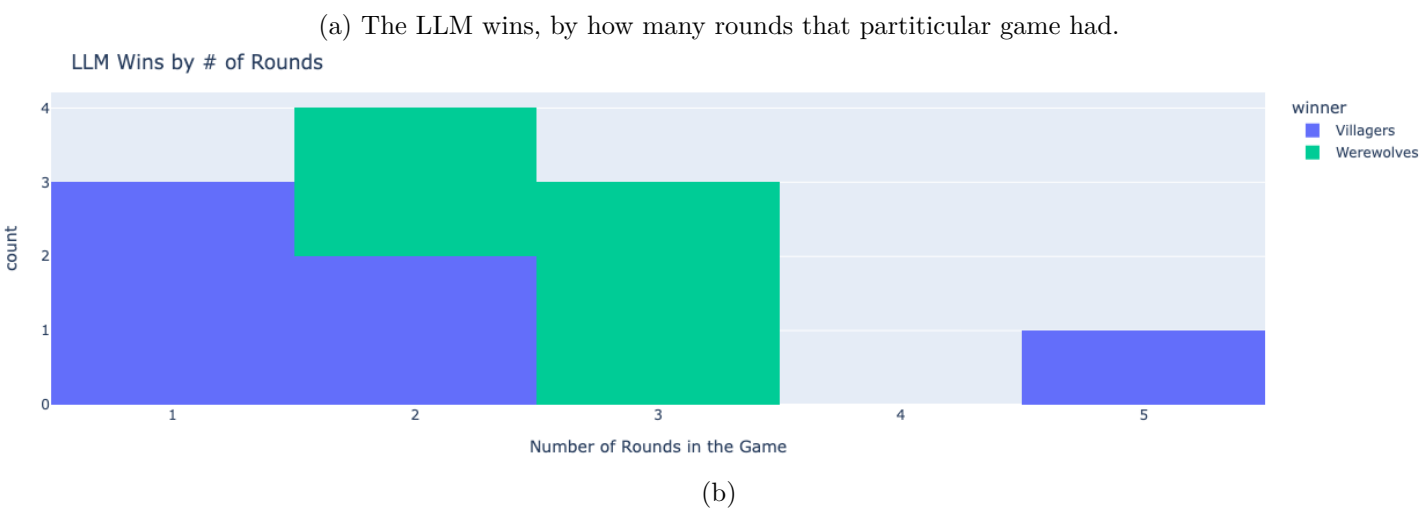


Figure 1

Source: [Werewolf Among Us: Human vs LLM Analysis](#)

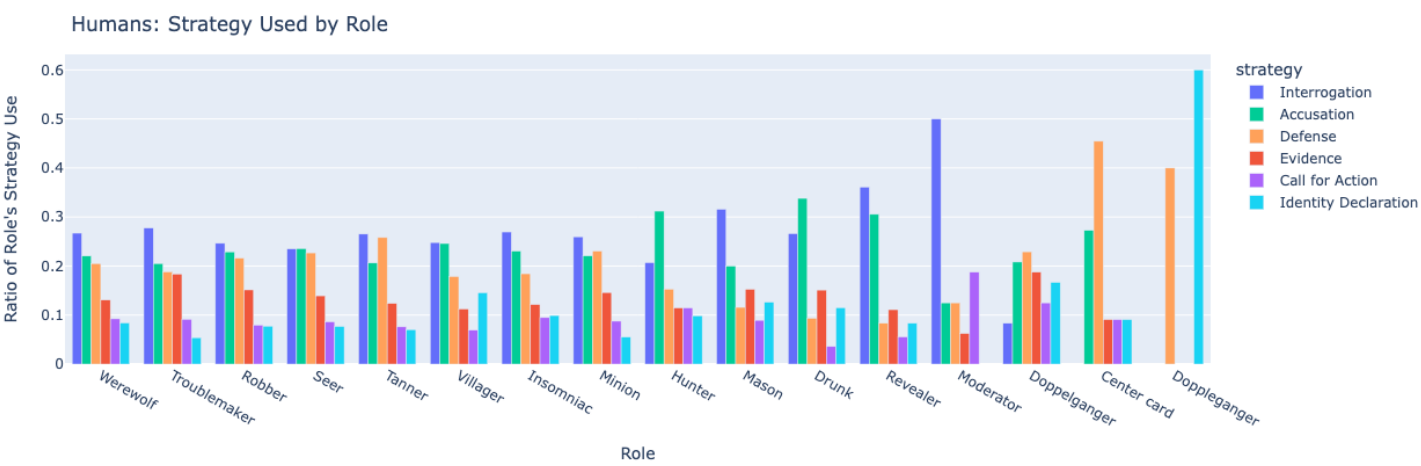


Figure 2

Source: [Werewolf Among Us: Human vs LLM Analysis](#)

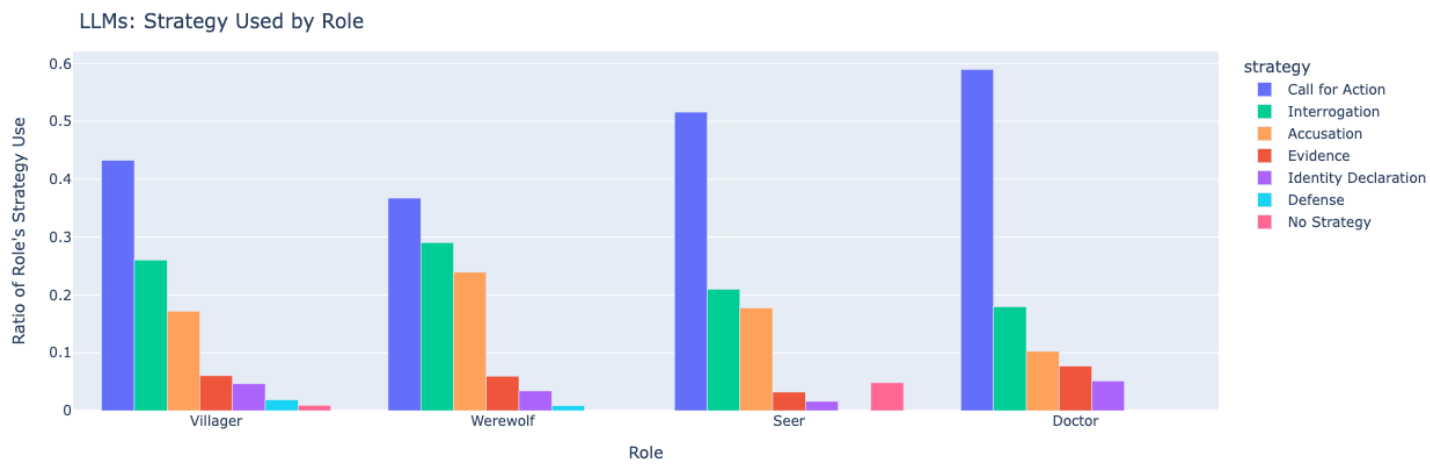


Figure 3

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 Alkhoul, 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, Yinyao, 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.
- Wikipedia contributors. 2024. “Mafia (Party Game).” [https://en.wikipedia.org/wiki/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