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

Bhavana Jonnalagadda Riley Jones

2025-05-06

Abstract TODO

Table of contents

Introduction **4**
Related Work . . . 4
Multi-Agent LLMs . . . 4
LLMs and Werewolf . . . 5

Methods **6**
Data . . . 6
Werewolf Among Us Human Dataset . . . 6
Werewolf Arena (LLM Dataset) . . . 6
Analysis . . . 6

Results **7**

Discussion and Conclusion **9**
Limitations . . . 9
Future Work . . . 9
Summary . . . 9

References **10**

Project Contributions **11**

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 have limited information, hidden identities, and must convince others while trying to figure out who is lying. These challenges closely match real life situations involving trust, negotiation, and manipulation.
- We wanted to compare how humans and large language models (LLMs) handle these situations. To do this, we used a recent human dataset and constructed one from an LLM-based simulator:
  - *Werewolf Among Us* (Lai et al. 2022), a collection of real human gameplay annotated with persuasion strategies,
  - *Werewolf Arena* (Bailis, Friedhoff, and Chen 2024), a simulated environment where LLM agents play the game autonomously.
  - While both studies show Werewolf generates complex strategic language, neither compares human and LLM behavior directly.
- Our project addresses this gap. We analyzed transcripts from both datasets, matched them by role, round, and persuasion strategy, and compared how humans and LLMs lie, persuade, and detect deception.
- By annotating utterances with the same set of persuasive strategies, we clearly show how synthetic agents differ from or resemble humans when navigating deception in 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 exisiting 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), which contains annotated dialogues from over 150 real games of One Night Werewolf and Avalon. These games differ from classic Werewolf because they have only one round of discussion and voting, do not eliminate players during the game, and include many specialized roles beyond Villager and Werewolf. The dataset provides detailed annotations of persuasion strategies for each utterance, such as accusations, defenses, and identity claims. We specifically used the textual transcriptions and their strategy annotations for our comparisons.

Werewolf Arena (LLM Dataset)

The *Werewolf Arena* dataset (Bailis, Friedhoff, and Chen 2024) features simulated classic Werewolf games played by LLM agents without human intervention. Unlike the one-round human games, these simulations involve multiple rounds alternating between night (secret actions) and day (open discussion). Each LLM agent is assigned a role (Villager, Werewolf, Seer, Doctor) and receives context-specific prompts that guide their strategic interactions.

We ran simulations using five LLM models: GPT-4o, GPT-4.1, GPT-4o-mini, DeepSeek-Chat, and DeepSeek-Reasoner. We experimented with two configurations:

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

We chose these configurations to give villagers more time to coordinate, extending the dialogue to better observe persuasive behaviors. After running the simulations, we manually annotated the LLM-generated dialogue using the same persuasion categories as in the human dataset.

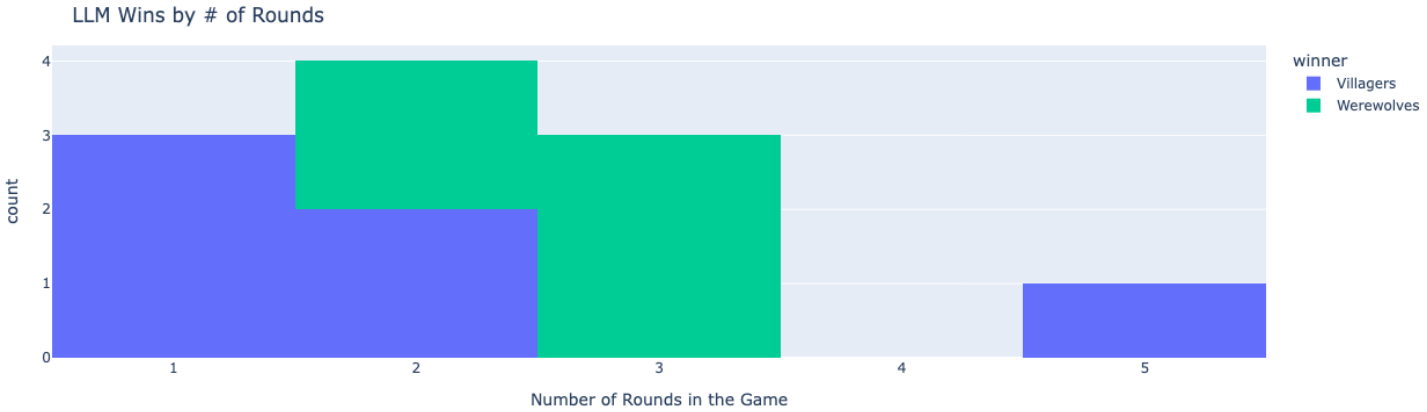
Analysis

We standardized annotations across both datasets so they could be directly compared. Our analyses focused on frequency distributions of persuasion strategies, comparisons by role (villager versus werewolf), and differences in how humans and LLMs applied these strategies.

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

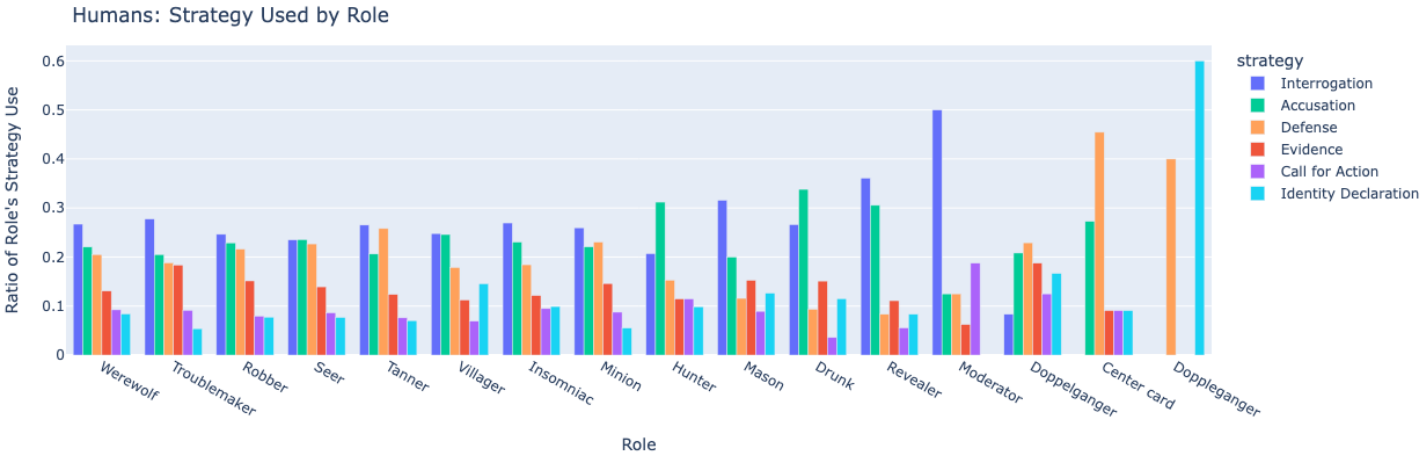


Figure 2

Source: Werewolf Among Us: Human vs LLM Analysis

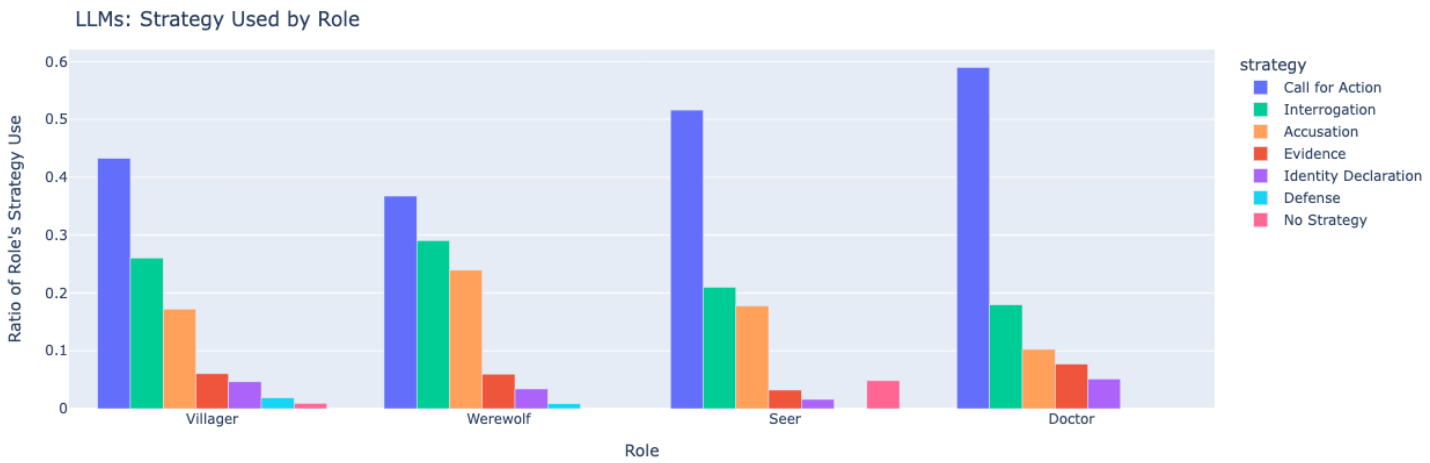


Figure 3

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.

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