

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

Bhavana Jonnalagadda

Riley Jones

2025-05-01

Abstract TODO

Table of contents

Introduction	4
Related Work	4
Multi-Agent LLMs	4
LLMs and Werewolf	4
Methods	5
Data	5
Werewolf Among Us Human Dataset	5
Werewolf Arena	5
Analysis	5
Results	6
Discussion and Conclusion	8
Limitations	8
Future Work	8
Summary	8
References	9
Project Contributions	10

Introduction

- A description of the problem and its significance
- How do LLMs function in a multi-agent environment?
 - Each has limited information
 - Approximately the same ability, traits, skills
- We use LLMs to simulate whether synthetic agents can participate in complex, adversarial group dynamics
- Werewolf is a good candidate for testing multi agent systems of cooperation and secrecy
 - Need citation
 - The game tests adaptive reasoning, strategic alignment, and collective threat detection under special conditions

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

- Human dataset description ([Lai et al. 2022](#))
- Is specifically for a form of one-night werewolf
 - Describe key differences
- Used specifically for the text available
 - and annotations of persuasion strategy on the text

Werewolf Arena

- ([Bailis, Friedhoff, and Chen 2024](#))
- Discuss the framework, how it works, prompts, etc
- Discuss what types of runs we did
- Discuss the data included in output
- Talk about how we had to annotate the LLM speech with persuasion strategies ourselves

Analysis

- Formatted data to match, performed various comparisons

Results

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

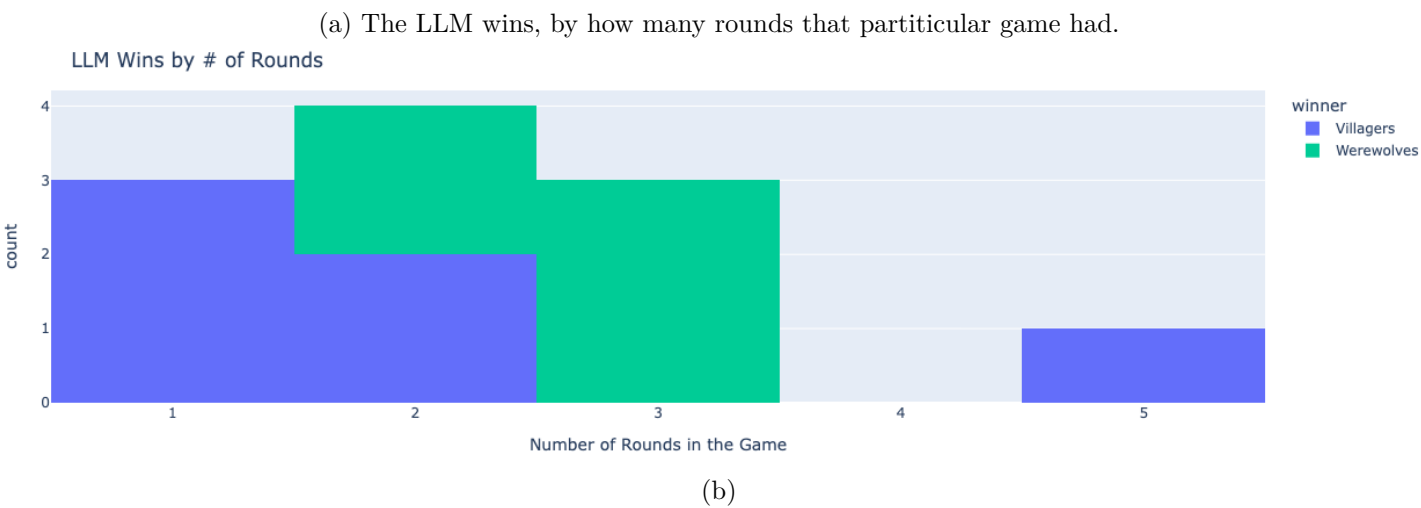


Figure 1

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

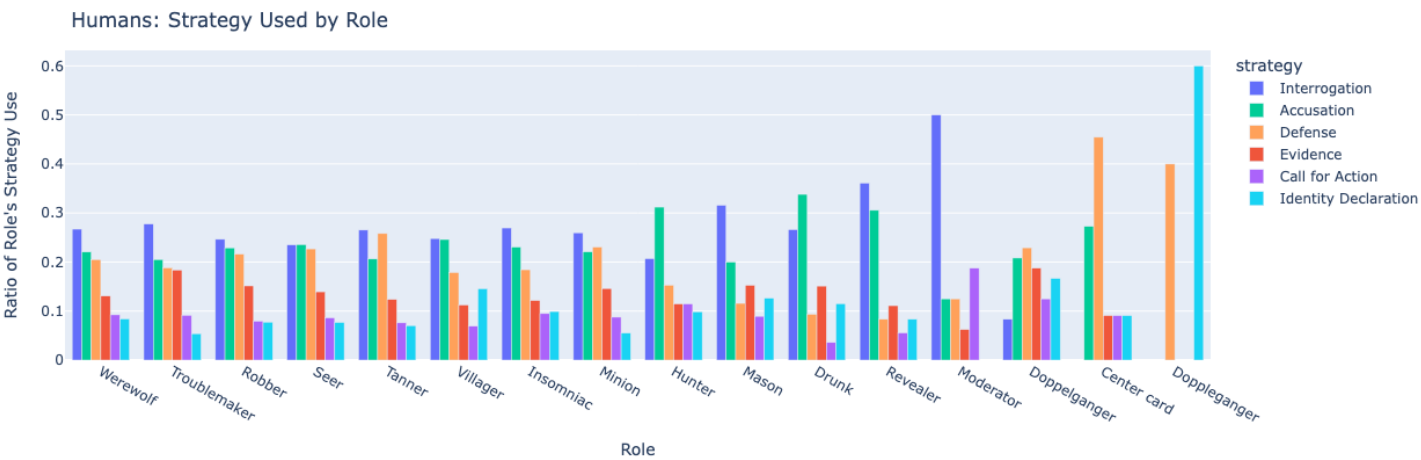


Figure 2

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

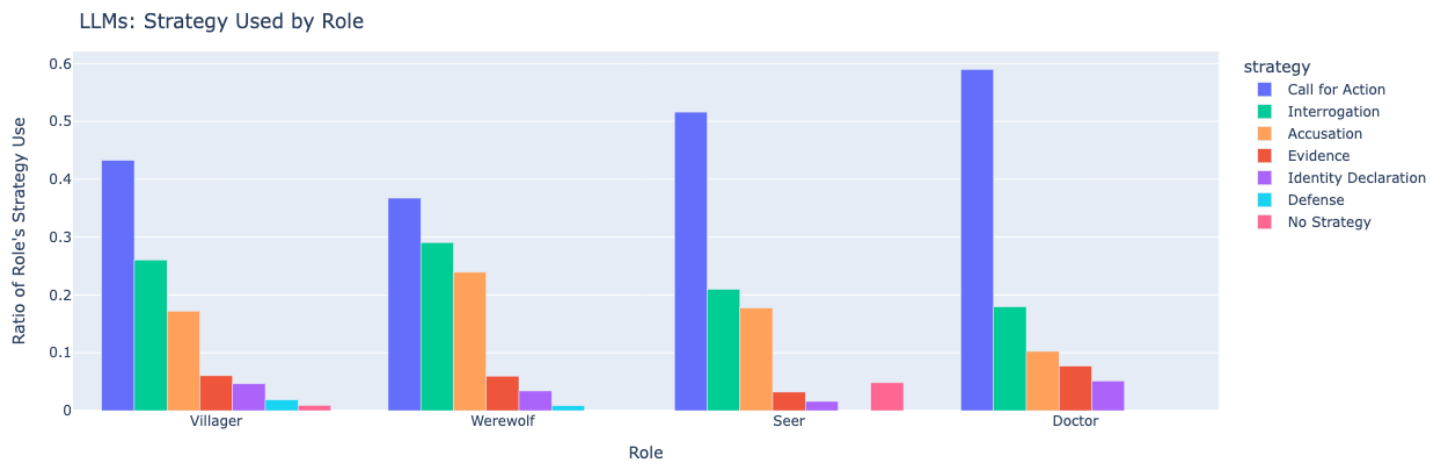


Figure 3

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 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.
- 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