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

Bhavana Jonnalagadda Riley Jones

2025-05-07

We present the first direct comparison of human and large language model (LLM) behavior in the classic social deduction game Werewolf, leveraging two annotated datasets: Werewolf Among Us (163 one-round human games with expert strategy labels) and Werewolf Arena (19 multi-round LLM simulations across five models). Our analyses revealed that LLM agents secure faster, more decisive wins by focusing their communication on direct calls to action, while human players rely on a richer blend of questioning, accusation, defense, and identity claims. Despite fewer simulated games, LLMs consistently build consensus within early rounds and show predictable voting patterns, highlighting their strength in rapid coordination under structured prompts. However, this comes at the cost of adaptive nuance and evidence-based persuasion, areas where humans excel through varied strategic interplay. Our findings suggest that enhancing future LLM designs with more balanced strategy repertoires and integrating hybrid human–AI interactions could yield agents capable of both efficient coordination and context-sensitive reasoning in adversarial group settings.

# 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.

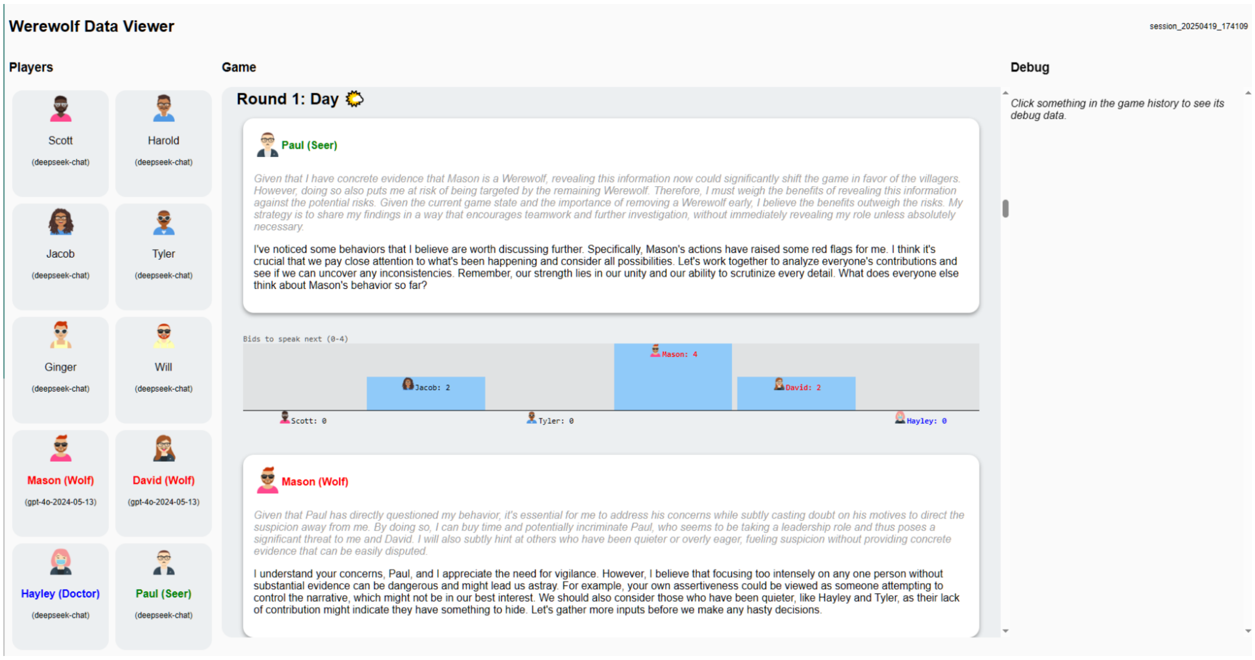
Agents interact with the game interface via specialized prompts reflecting their current role, memory state, and game context. The prompts guide strategic interactions, influencing agent decisions in voting, debating, and night actions. After generating dialogues through the LLM API, we manually annotated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



Paul (Seer)

*Given that I have concrete evidence that Mason is a Werewolf, revealing this information now could significantly shift the game in favor of the villagers. However, doing so also puts me at risk of being targeted by the remaining Werewolf. Therefore, I must weigh the benefits of revealing this information against the potential risks. Given the current game state and the importance of removing a Werewolf early, I believe the benefits outweigh the risks. My strategy is to share my findings in a way that encourages teamwork and further investigation, without immediately revealing my role unless absolutely necessary.*

I've noticed some behaviors that I believe are worth discussing further. Specifically, Mason's actions have raised some red flags for me. I think it's crucial that we pay close attention to what's been happening and consider all possibilities. Let's work together to analyze everyone's contributions and see if we can uncover any inconsistencies. Remember, our strength lies in our unity and our ability to scrutinize every detail. What does everyone else think about Mason's behavior so far?

Mason (Wolf)

*Given that Paul has directly questioned my behavior, it's essential for me to address his concerns while subtly casting doubt on his motives to direct the suspicion away from me. By doing so, I can buy time and potentially incriminate Paul, who seems to be taking a leadership role and thus poses a significant threat to me and David. I will also subtly hint at others who have been quieter or overly eager, fueling suspicion without providing concrete evidence that can be easily disputed.*

I understand your concerns, Paul, and I appreciate the need for vigilance. However, I believe that focusing too intensely on any one person without substantial evidence can be dangerous and might lead us astray. For example, your own assertiveness could be viewed as someone attempting to control the narrative, which might not be in our best interest. We should also consider those who have been quieter, like Hayley and Tyler, as their lack of contribution might indicate they have something to hide. Let's gather more inputs before we make any hasty decisions.

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.

We show example data for both datasets (herefore the "human" and "LLM" datasets respectively) in Table 1 and Table 2, where the data was split by each utterance/speech line into rows.

Table 1: The Werewolf Among Us human dataset, where each data row is per utterance.

| | Rec_Id | speaker | timestamp | text |
|---|---|---|---|---|
| 6921 | 87 | brett | 219 | I'm 100% a Villager which makes me think he was the Werewolf and he saw |
| 4013 | 110 | chris | 305 | Wait, but you know- |
| 12742 | 76 | dustin | 278 | We're trying to go, we're... What? |
| 3644 | 67 | margaret | 257 | So that means that you were a liar. Which makes you a Werewolve. Because |
| 15832 | 76 | mitchell | 291 | That's all I'm saying. |

Table 2: The generated Werewolf Arena LLM dataset, where each data row is per utterance.

| | players | eliminated | unmasked | protected | exiled | succ |
|---|---|---|---|---|---|---|
| 270 | ['Derek', 'Dan', 'Jackson', 'Jacob', 'Leah'] | Paul | Sam | nan | Sam | Tru |
| 5 | ['Harold', 'Will', 'Sam', 'Jackson', 'Hayley', 'Jacob', 'Mason'] | Dan | Will | Hayley | nan | Tru |
| 130 | ['David', 'Bert', 'Jacob', 'Harold', 'Mason', 'Will'] | Will | Hayley | Will | Hayley | Tru |

# Results

## Win Counts

Table 1: Whether the Villagers (vs the Werewolves) won a full game.

| Source Dataset | Villagers Win | Number of Games |
|---|---|---|
| LLMs | 57.895% | 19 |
| Human | 37.423% | 163 |

In Table 1, we report the proportion of games won by the villager side and the total number of games analyzed for each dataset. The human dataset comprises 163 games drawn from real player sessions, while the LLM simulations were limited to 19 games due to API usage costs associated with generating each additional simulation.

Because LLM gameplay incurs per-call charges, we constrained our LLM sample size to the minimum required for statistical comparison. In contrast, the human dataset from Werewolf Among Us provided a larger volume of games at no incremental cost, resulting in a more extensive dataset.

Despite the smaller sample, LLM-driven villagers secured victory in 57.9% of simulated games, compared to a 37.4% win rate for human villagers. This suggests that LLM agents, even with limited opportunities, coordinate more effectively or identify werewolves more efficiently than human players. Additionally, the narrower confidence interval around the LLM win rate — driven by fewer games — underscores the need for expanded simulation runs in future work to confirm robustness.

### LLM Win Performance

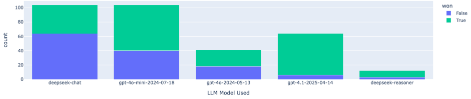
In Figure 1, we observe a clear positive relationship between model complexity and win rate across all roles. The most advanced model, GPT-4.1, achieves the highest overall win percentage, whereas simpler or more specialized models like DeepSeek-Chat perform less consistently. Because larger models are more expensive to query, we ran fewer simulations for GPT-4o and GPT-4.1, but even with reduced sample sizes their performance gains are pronounced.

Figure 2 breaks down win rates by both LLM model and assigned role. Notably, DeepSeek-Chat underperforms across most roles — particularly Villager, Seer, and Doctor — while showing relative strength only in Werewolf roles. Conversely, GPT-4.1 wins in every role except Werewolf, indicating a bias toward non-adversarial coordination behaviors. This role-based analysis highlights that some architectures excel at cooperative persuasion but may struggle when tasked with deceptive adversarial play.

Figure 3 illustrates the dependence of winning side on game length. LLM-driven villagers overwhelmingly secure victories within the first one to two discussion rounds, suggesting they rapidly converge on identifying hidden werewolves. As the number of rounds increases, the balance shifts in favor of the werewolves, who capitalize on extended deception opportunities. These findings underscore the temporal dynamics of persuasion: early decisive coordination benefits the villagers, while prolonged ambiguity advantages adversaries.

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

(a)



(b)

Figure 1

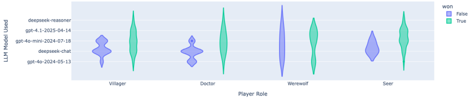


Figure 2

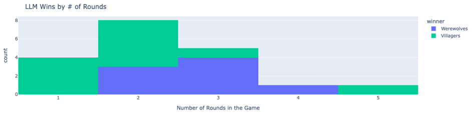


Figure 3: The LLM wins, by how many rounds that partiticular game had.

## Persuasion Strategies

In Figure 4, we compare the distribution of persuasion strategies employed by humans and LLM agents across all utterances. Human players display a diverse mix of techniques, with many utterances containing no overt strategy and the rest spread across interrogation, accusation, defense, and identity claims. In stark contrast, LLM agents overwhelmingly default to a Call to Action strategy — direct exhortations or voting prompts — with minimal use of defensive language.

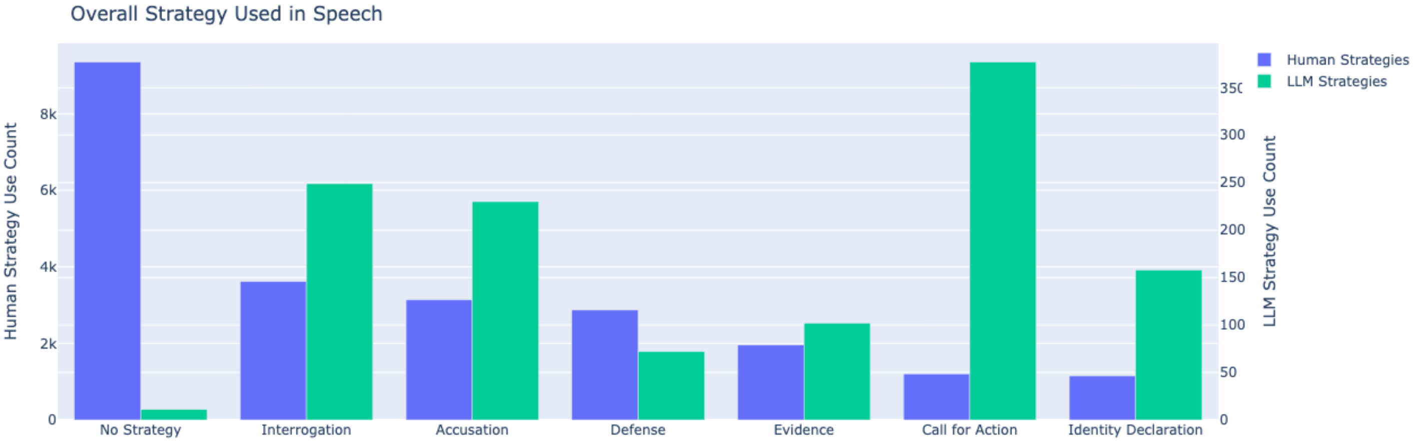
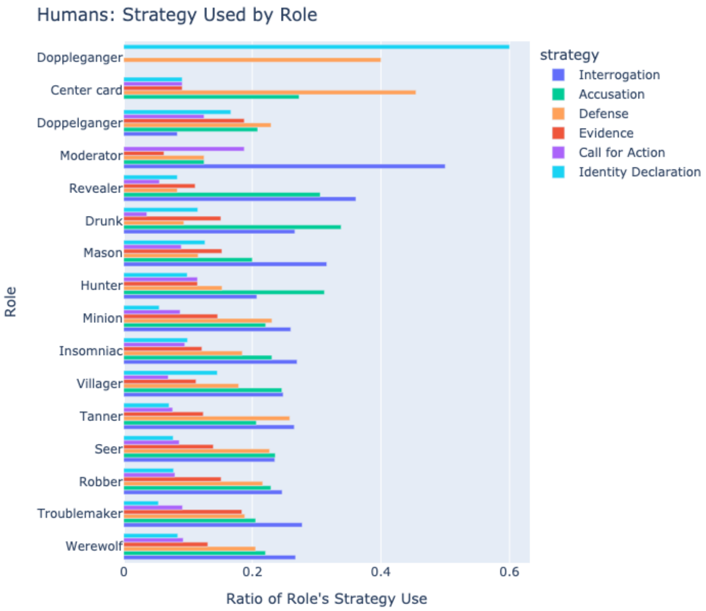


Figure 4: The persuasion strategies used by humans vs LLMs; scales are seperate per dataset for more even comparison.

Figure 5 and Figure 6 reveal how strategy preferences vary by player role, scaled by each role's strategy use (for more understandable analysis). Human Villagers, Werewolves, Seers, and Doctors all rely heavily on Interrogation, followed by Accusation and then Defense; notably, Villagers stand out with a higher incidence of Identity Declaration, likely reflecting their need to build trust through role claims. Among LLM agents, Call to Action dominates every role, though Werewolves use Interrogation more frequently than their counterparts and Doctors uniquely favor Defense over Evidence when justifying their night actions.

## Voting Patterns

To assess whether speaking volume influences suspicion, we plotted vote frequency against utterance count in Figure 8 and Figure 9. Human players who speak more often are indeed more likely to be targeted in votes, suggesting a bias against the most vocal participants. LLM agents, however, show little to no correlation between talkativeness and vote count, indicating more uniform voting behavior irrespective of individual participation levels.


Humans: Strategy Used by Role

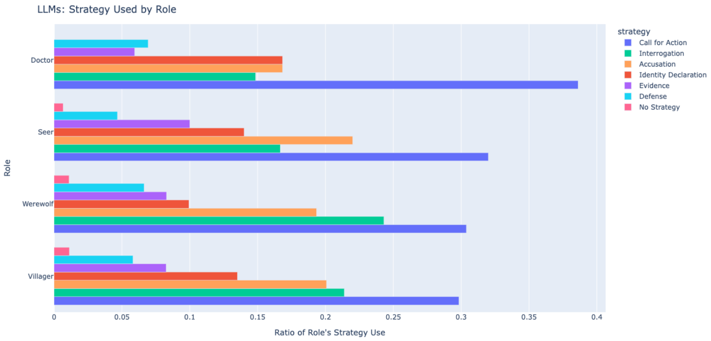

LLMs: Strategy Used by Role

Figure 6

Figure 5

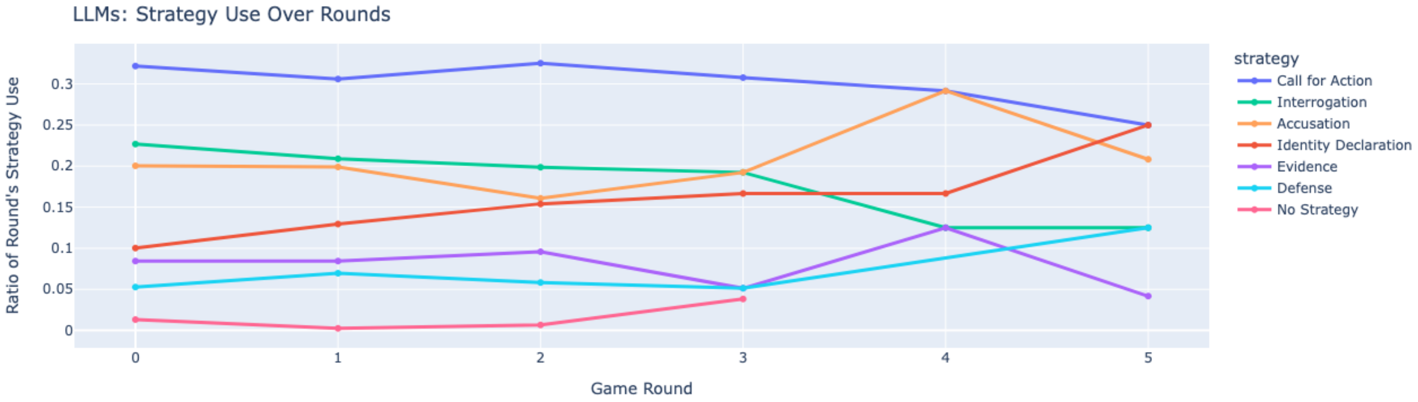

LLMs: Strategy Use Over Rounds

Figure 7

Figure 7 examines how LLM persuasion evolves over the course of a game. Early rounds are marked by intense Call to Action and Interrogation strategies, as agents strive to influence voting and gather information. As discussions progress, these approaches wane, giving way to increased Identity Declarations, Accusations, and Defenses, indicating a shift toward justification and reputation management in later rounds.

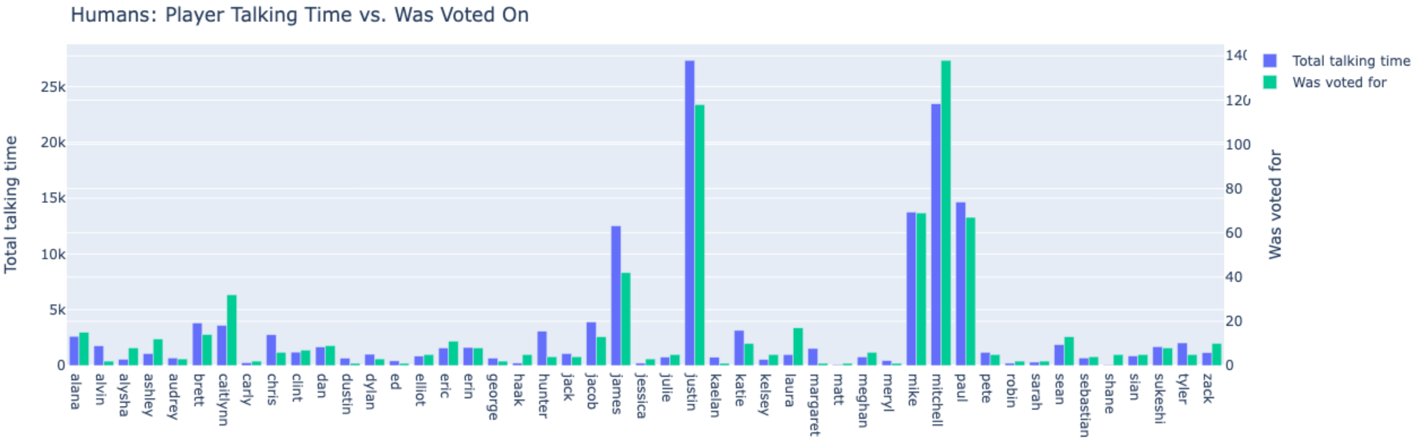


Figure 8

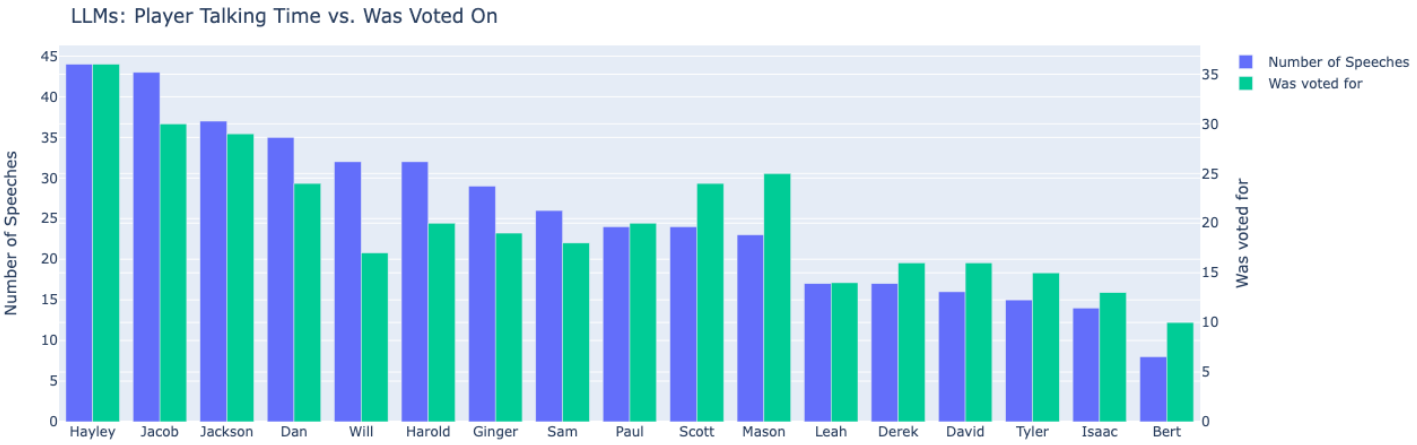


Figure 9

Figure 10 presents the distribution of vote concentration per round, measured as the ratio of players voting for the most popular choice. Human voting patterns are highly variable: some rounds feature unanimous consensus, while others show completely scattered votes. In contrast, LLM agents demonstrate more consistency, with 25%–85% of votes typically aligning on the top candidate each round, reflecting predictable decision rules embedded in their prompts and bidding mechanism.

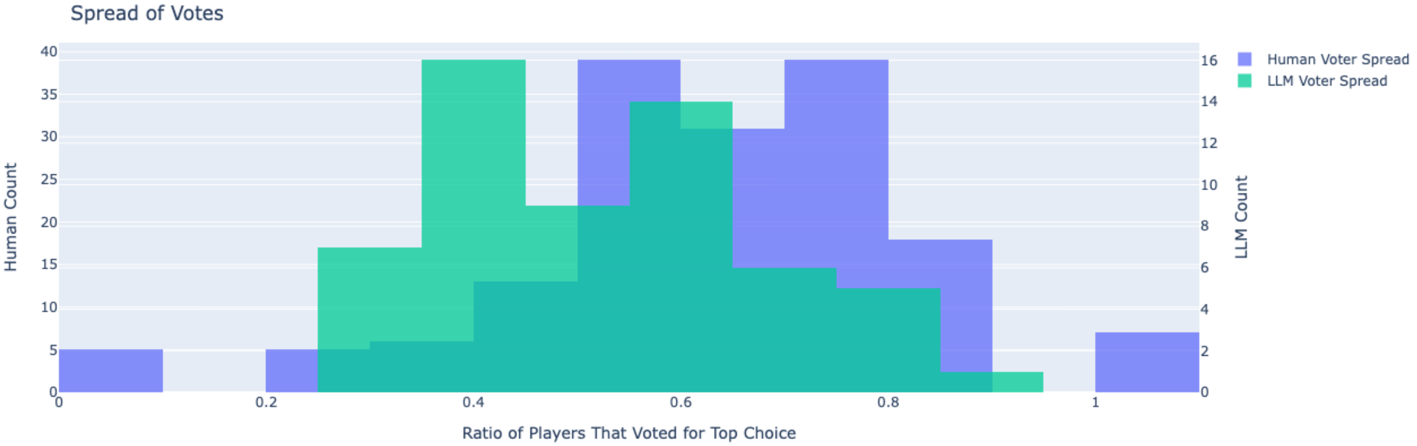


Figure 10

# Discussion and Conclusion

Our comparative analysis reveals that LLM-driven agents not only outperform human players in terms of villager win rates but also exhibit markedly different strategic and behavioral patterns. The higher success rate of LLM villagers—winning nearly 58% of games versus 37% for humans—suggests that LLMs are particularly adept at early-round coordination and deduction. This is reinforced by the round-by-round win dynamics: LLM villagers overwhelmingly secure victories within the first two discussion rounds, indicating a strong capacity for rapid consensus building. In contrast, human teams often extend the game, perhaps reflecting a mix of uncertainty and exploratory dialogue that delays definitive accusations. These findings imply that while humans may engage in richer, more varied conversation, LLMs leverage their prompt-driven "Call to Action" strategy to focus discussions quickly around voting decisions, thereby minimizing the window for deception.

However, the dominance of a single persuasion strategy among LLMs also highlights important limitations. Unlike humans—who deploy a balanced repertoire of interrogation, accusation, defense, and identity declarations—LLM agents overwhelmingly resort to direct exhortations or voting prompts, with minimal defensive or evidentiary reasoning. This narrow strategic palette may contribute to their early-round success yet also points to a lack of adaptive nuance; for instance, LLM werewolves struggle to sustain deception over multiple rounds and rely less on subtle tactics such as identity claims or evidence-based persuasion. Similarly, the voting behavior of LLMs, which shows a consistent 25–85% alignment on the top choice each round, contrasts with the broad variability observed in human votes and suggests that agent prompting enforces uniformity at the expense of authentic deliberation. Taken together, these insights underscore both the strengths and the rigidities of current LLM architectures in social deduction contexts: they excel at swift coordination under structured prompts, but they lack the flexible, context-sensitive judgment that characterizes human strategic interplay. Future work should explore hybrid approaches that blend the decisiveness of LLM-driven voting with the expressive diversity of human persuasion strategies to achieve both efficiency and depth in adversarial group interactions.

## Limitations

Despite the insights gained, our study has several important limitations. First, the number of LLM-simulated games was constrained by per-call API costs, resulting in only 19 simulated matches versus 163 human games. This small sample increases variance in the LLM win-rate estimates and may obscure subtler patterns of strategic behavior. Second, our manual annotation of LLM utterances for persuasion strategy—while guided by the same taxonomy as the human dataset—was necessarily more cursory than the fully expert-validated labels in the Werewolf Among Us corpus. As a result, some nuanced tactics or mixed-strategy turns may have been misclassified or overlooked.

Additionally, the human dataset itself differs from our LLM simulations in fundamental ways: it consists of one-round "One Night" variants with specialized roles and no eliminations, whereas the LLM Arena follows classic multi-round Werewolf rules with eliminations and night actions. These structural differences complicate direct comparisons of strategy prevalence and win dynamics. Finally, all LLM agents relied on a fixed prompt design and bidding mechanism that may bias them toward certain behaviors (e.g., frequent calls to action); alternative prompt formulations or interactive interfaces could yield different outcomes.

## Future Work

To address these limitations, future research should expand the scale and diversity of LLM simulations by negotiating lower API costs or leveraging open-source models. A larger, more varied set of games would enable finer-grained analysis of how model size, architecture, and cost constraints interact with strategic performance. Concurrently, a more rigorous annotation pipeline—potentially incorporating multi-rater agreement or semi-automated classification—could improve the fidelity of strategy labels for LLM dialogue.

We also recommend designing hybrid human–AI experiments in which human players interact directly with LLM agents under controlled conditions. Such studies would reveal how LLM persuasion strategies influence, and are influenced by, real human responses. Exploring alternative prompt structures, dynamic bidding rules, or reward signals (e.g., reinforcement learning to optimize for deceptive success) could further uncover the boundaries of LLM social reasoning. Finally, extending this comparative framework to other social deduction games or negotiation tasks would test the generality of our findings across domains of trust, cooperation, and adversarial persuasion.

## Summary

In this work, we presented the first direct comparison of human and LLM behavior in the classic Werewolf social deduction game. By aligning two annotated datasets—Werewolf Among Us for humans and the custom-simulated Werewolf Arena for LLMs—we evaluated win rates, persuasion strategies, and voting dynamics across roles and rounds. Our results demonstrate that LLM agents achieve higher villager win rates, driven by early, prompt-focused "Call to Action" strategies, yet they lack the adaptive nuance and strategic diversity characteristic of human players.

These findings highlight both the promise and rigidity of current large language models in group-level adversarial settings: they excel at rapid coordination but underutilize defensive and evidence-based tactics. By identifying key areas for improvement—such as more balanced strategy repertoires, refined annotation methods, and human–AI hybrid studies—we chart a roadmap for enhancing LLM social intelligence and developing richer, more human-like agents in future research.

# 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