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

Bhavana Jonnalagadda Riley Jones

2025-04-23

This final project for CSCI-5423 explores the modeling of persuasion behaviors in the text-based social deduction game Werewolf. We leverage the multimodal "Werewolf Among Us" dataset (Lai, 2022) and evaluate several large language models on their ability to persuade, deceive, and cooperate within game dialogues. Through a combination of feature engineering, sequence modeling, and reinforcement learning agents, we compare performance against baseline classifiers and analyze key linguistic strategies.

# Table of contents

# Introduction

Introduce the motivation for studying persuasion in social deduction games. Describe the Werewolf game mechanics and why it offers a rich testbed for multimodal persuasion analysis.

# Related Work

Review prior datasets and frameworks:

- Werewolf Among Us: dataset and original paper (Lai, 2022)
- PersuasionGames repository and baseline models
- AmongAgents and evaluation of LLMs in interactive text-based games (Chi, 2024)
- RL approaches to strategic play in Werewolf (Xu et al., 2023)
- Werewolf Arena framework by Google for LLM evaluation

# Data and Preprocessing

Detail how we load and preprocess the HuggingFace Werewolf-Among-Us dataset. Include data splits, feature extraction, and any augmentation steps.

# Methods

Describe the modeling approaches:

- Baseline classifiers (e.g., logistic regression, SVM)
- Sequence models (e.g., LSTM, Transformer)
- RL-based agent setup and reward definitions

# Experiments and Evaluation

Outline experimental setups:

- Persuasion classification tasks
- Game simulation with LLM agents
- Metrics for persuasion success, deception detection, and cooperative play

# Results

Present quantitative results in tables and figures. Analyze which models and features best capture persuasion strategies.

# Discussion

Interpret findings, discuss limitations, and propose future work.

# Conclusion

Summarize contributions and insights from the project.

# References

Lai, B. (2022). Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games. arXiv:2212.08279.

Chi, Y. (2024). AmongAgents: Evaluating Large Language Models in the Interactive Text-Based Social Deduction Game. arXiv:2407.16521v2.

Xu, Z., et al. (2023). Language Agents with Reinforcement Learning for Strategic Play in the Werewolf Game. arXiv:2310.18940v3.

Blanchard, T., et al. (2024). Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction. arXiv:2407.13943.