CS 557: Grand Challenge

Frederic Guintu Allison Ng William Hao
Jose Izaguirre

December 1, 2025

1 Introduction

Our team proposes a project titled *Adversarial Prompting via Social-Deduction Search*, inspired by the NeurIPS 2025 MindGames Arena Hub social-deduction environment, specifically the Mafia setting [1]. This environment combines hidden information, strategic dialogue, and adversarial reasoning, ideal for testing the robustness of LLM prompts in competitive multi-agent settings. The problem we aim to solve is that typical prompting strategies often break under adversarial pressure: models may leak role information, contradict prior statements, or struggle to deceive and deduce effectively. Our baseline will replicate standard LLM-based Mafia simulations using static role prompts and simple reasoning chains.

2 Proposed Approach

To improve on this, we will build prompting strategies that emulate known optimal deception and reasoning tactics from both AI and human play. From AI, we draw on identity-detection reinforcement learning (IDRL), which trains agents to infer hidden roles while balancing collaboration and deception [2]. We plan to adapt this by maintaining a running belief state for each LLM agent, updated after each interaction using manually defined heuristics (similar to scores we defined in the pacman projects e.g., suspicion score: increases if a player contradicts past statements or avoids discussion). These belief updates will then act as an expectimax node for our agents next action, influencing the agent's conversational goals (e.g., accusing, defending, withholding information). From human Mafia play, we use a modified version of the "Random+All-In" strategy, which mixes passive bluffing with sudden coordinated moves [3]. In our case, given the assumption that other agents will be playing optimally, we would introduce a random aspect in our agent's play to disrupt the strategies of other agents. Our agents will lie sparingly but strategically, maintain consistent falsehoods, and avoid suspicious behavior guided by effective Mafia gameplay theory.

3 Model Evaluation and Assessment

We will evaluate our approach using the MindGames Mafia setup and other available multi-agent social deduction environments. Metrics include win rate, hidden role leakage, consistency under pressure, and bluff effectiveness. Transcripts will be analyzed for how well the LLM maintains character, successfully deceives, or detects opponents. We expect our approach to outperform the baseline in maintaining role secrecy and applying deception tactics.

References

[1] MindGames Arena Hub. NeurIPS 2025 Theory-of-Mind Challenge.

[2] Shijie Han, Siyuan Li, Bo An, Wei Zhao, Peng Liu (2023). Classifying Ambiguous Identities in Hidden-Role Stochastic Games with Multi-Agent Reinforcement Learning

[3] Shitong Wang. (2024). Optimal Strategy in Werewolf Game: A Game Theoretic Perspective