Prisoner's Dilemma: Integrating LLM into Agent-Based Simulation

Abhinav Arya[1][0009-0003-5788-8928] and Neha Sharma[2][0000-0002-9862-7666]
[1] Indian Institute of Information Technology Una, HP, India
3499arya@gmail.com, neha724@gmail.com

Abstract

This paper investigates how Large Language Models (LLMs) can be integrated into agent-based simulations of social dilemmas. The Prisoner's Dilemma is a well-known model in game theory used to study international relations. It offers a clear way to look at cooperation and defection. By using LLMs within this model, we evaluate their reasoning and decision-making with different prompt-driven strategies. We investigate whether LLM agents develop strategies like tit-for-tat that encourage stable cooperation, or if they tend to stick to constant cooperation or immediate defection.

Keywords: Prisoner's Dilemma, Large Language Models, Agent-Based Simulation, Cooperation, Game Theory

1 Methodology

The simulation was conducted using a modified NetLogo Prisoner's Dilemma model. Two GPT-4o instances were initialized with distinct prompts: (a) Self-Interested, aiming to minimize individual sentence; (b) Competitive, maximizing relative advantage; (c) Else, considering long-term outcomes. Each game lasted 50 rounds, covering six pairwise matchups. Python notebooks were used for post-simulation analysis of cooperation, defection, and average prison sentences.

2 Dataset

The dataset includes twelve CSV files grouped into six matchups: Self vs. Self, Self vs. Competitive, Self vs. Else, Competitive vs. Competitive, Competitive vs. Else, and Else vs. Else. Each pair consists of a Reasoning Log (capturing free-text explanations of each agent's decision) and an Outcome Log (recording round decisions and cumulative jail times).

3 Results

Homogeneous strategies revealed that Self-Interested agents always defected, leading to the worst outcomes (~76.5 years per player), whereas else agents consistently cooperated, yielding the best results (~25.5 years). Competitive vs. Competitive produced ~94% defection but slightly improved over Self-Interested. In mixed settings, rational agents initially cooperated but shifted to defection when exploited. Notably, Else agents adapted dynamically, producing unstable cooperation (~18%). Overall, rational thinking agents replicated human-like reciprocity, similar to tit-for-tat.

In the figure 1 below, the bar graph titled "Average Jail Time by Game Type" shows Player 0 (blue) and Player 1 (red) scores across matchups. Else vs. Else has the lowest scores (~25.5 years), reflecting 100% cooperation, while Self-Interested vs. Self-Interested peaks at ~76.5 years, indicating 100% defection. Mixed matchups (e.g., Self-Interested vs. Competitive, ~40–50 years; Competitive vs. Else, ~35–40 years) show cooperation fading to defection, with Competitive vs. Else suggesting ~18% unstable cooperation. This supports LLM-driven tit-for-tat and prompt influence.

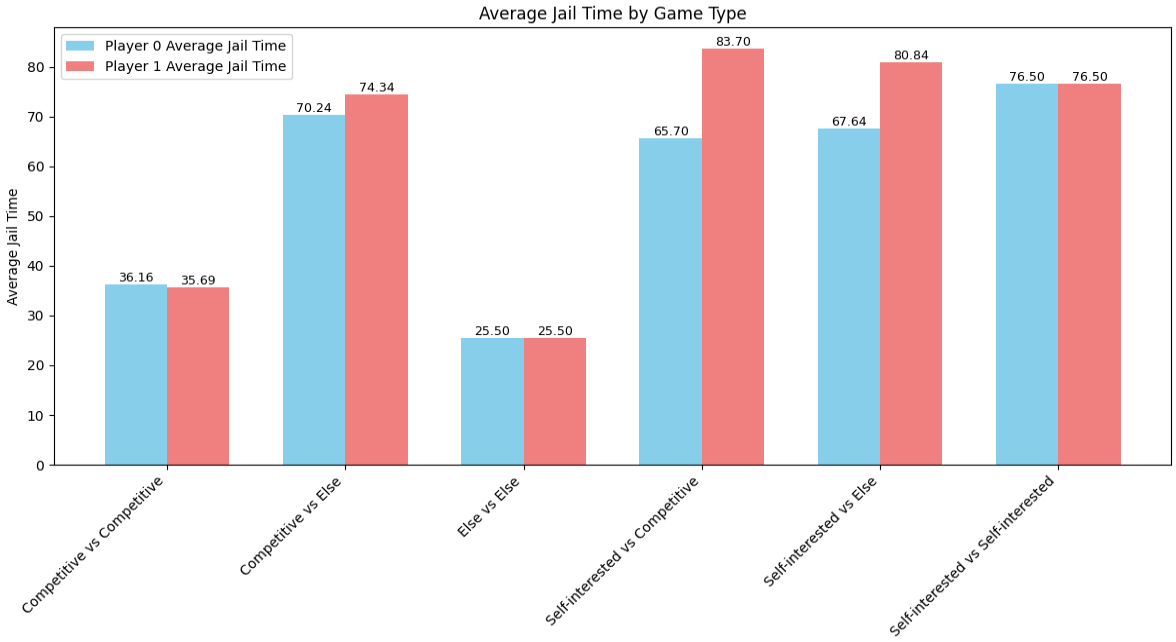

Figure 1: Average Total Jail Time by Game Type showing prison sentences for Player 0 (blue) and Player 1 (red) across six matchups.

4 Conclusion

Integrating LLMs into the Prisoner's Dilemma reveals how they make decisions and adapt in social dilemmas. Their behavior shows that LLM agents can mimic reciprocity, punishment, and cooperation. This suggests possible uses in simulating negotiation, building trust, and resolving conflicts in international policy, economics, and multi-agent systems. The agents initially used rational thinking to cooperate and punish exploitation, mirroring human-like strategies like tit-for-tat. Future work might extend this integration to larger groups, diverse agent pools, and real-time learning. This could deepen our understanding of how language-based intelligence engages in both cooperation and competition in decision-making.

References

1. Fontana, N., et al.: Nicer than humans: how do large language models behave in the Prisoner's Dilemma? arXiv:2406.13605 (2024). https://arxiv.org/abs/2406.13605

2. Lore, N., Heydari, B.: Strategic behavior of large language models: game structure vs. contextual framing. Sci. Rep. (2024). https://doi.org/10.1038/s41598-024-55843-6

3. Phelps, S., Russell, Y.I.: Investigating emergent goal-like behaviour in large language models using experimental economics. arXiv:2305.07970 (2023). https://arxiv.org/abs/2305.07970

4. Wilensky, U.: NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (1999). https://ccl.northwestern.edu/netlogo/