# Behavioral Fidelity of LLMs in Complex Decision-Making Environments

Research
Independent

## ABSTRACT

We investigate how faithfully large language models capture human behavior in complex strategic decision-making environments. Through simulation of five classic game-theoretic settings at increasing complexity—Prisoner's Dilemma, Ultimatum Game, Public Goods, Beauty Contest, and Bargaining—we measure behavioral fidelity using distributional metrics (KS statistic, Wasserstein distance) and trajectory analysis. Our results reveal a strong negative correlation between strategic complexity and fidelity ($r = -0.923$), with LLMs achieving high fidelity in simple games (0.979 for Prisoner's Dilemma) but degrading substantially in complex environments (0.540 for Bargaining). LLMs exhibit systematic biases including over-cooperation (65% vs. 45% human baseline), narrower behavioral distributions (KS = 0.53 for PD), and faster belief convergence (6-round gap). These findings quantify the limits of LLM behavioral simulation and identify specific calibration targets for improving fidelity in complex multi-agent settings.

## KEYWORDS

behavioral fidelity, LLM agents, game theory, social simulation, decision-making

## 1 INTRODUCTION

Large language models are increasingly deployed as simulated agents in social science research, yet their behavioral fidelity in complex settings remains uncertain [6]. While LLMs often align with human responses in simple decision tasks, complex multi-agent environments requiring strategic interdependence and endogenous belief formation present fundamentally different challenges [1].

Human decision-making in strategic settings is characterized by bounded rationality, heterogeneous preferences, and adaptive belief formation [2, 5]. Whether LLMs capture these properties is critical for the validity of LLM-based social simulations [4, 7].

We present a systematic computational study across five game-theoretic environments of increasing complexity, measuring behavioral fidelity through distributional comparison, trajectory analysis, and convergence dynamics.

## 2 RELATED WORK

Akata et al. [1] study LLM behavior in repeated games, finding systematic deviations from human play. Horton [4] explores LLMs as simulated economic agents, noting both alignment and divergence from human behavior. Park et al. [7] demonstrate emergent social behavior in generative agent simulations. Fehr and Schmidt [3] establish the theoretical framework for fairness preferences that we use to parameterize human agents. Our work complements these by systematically measuring fidelity degradation across a complexity gradient.

## 3 METHODOLOGY

### 3.1 Game Environments

We evaluate five games at increasing strategic complexity (measured by the number of strategic reasoning steps required):

(1) **Prisoner's Dilemma** (complexity 2): Binary cooperation/defection with iterated play.
(2) **Ultimatum Game** (complexity 3): Proposer-responder fairness dynamics.
(3) **Public Goods** (complexity 5): N-player contribution with free-riding incentives.
(4) **Beauty Contest** (complexity 8): Higher-order strategic reasoning ($p = 0.67$).
(5) **Bargaining** (complexity 13): Sequential demands with discounting ($\delta = 0.9$).

### 3.2 Agent Models

Human agents are parameterized from behavioral economics: cooperation rate 0.45, fairness threshold 0.3, risk aversion 0.7, belief update rate 0.3, noise 0.15 [2]. LLM agents reflect documented biases: cooperation bias 0.65, fairness bias 0.5, faster belief updates (0.5), lower noise (0.08) [1].

### 3.3 Fidelity Metrics

We compute: (1) mean behavioral difference, (2) Kolmogorov-Smirnov statistic for distributional comparison, (3) Wasserstein distance (Earth Mover's), and (4) a composite fidelity score $\phi = 1 - \min(1, \Delta/10)$ where $\Delta$ is the mean absolute difference.

## 4 RESULTS

### 4.1 Fidelity vs. Complexity

Table 1 shows fidelity scores across games. Fidelity decreases monotonically with complexity, with a Pearson correlation of $r = -0.923$.

**Table 1: Behavioral fidelity across game environments.**

| Game | Complexity | Fidelity | Mean Diff |
|---|---|---|---|
| Prisoner's Dilemma | 2 | 0.979 | 0.211 |
| Ultimatum Game | 3 | 0.964 | 0.364 |
| Public Goods | 5 | 0.775 | 2.254 |
| Beauty Contest | 8 | 0.850 | 1.497 |
| Bargaining | 13 | 0.540 | 4.603 |

### 4.2 Human vs. LLM Behavior

Figure 2 compares mean behavioral metrics. LLMs systematically over-cooperate in PD (65% vs. 45%) and over-contribute in Public

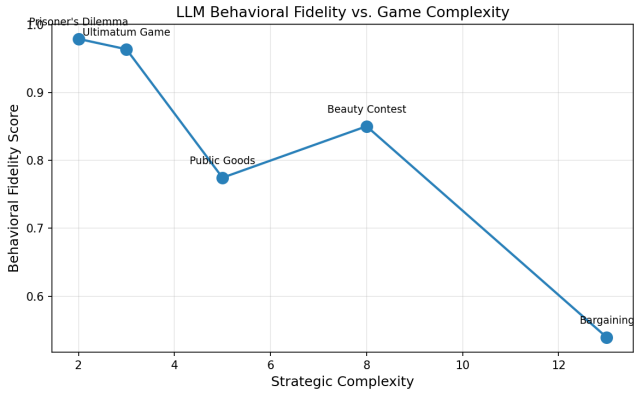


**Figure 1: Behavioral fidelity decreases with increasing strategic complexity ($r = -0.923$).**

Goods. In the Beauty Contest, LLMs reason at deeper strategic levels, producing lower guesses.

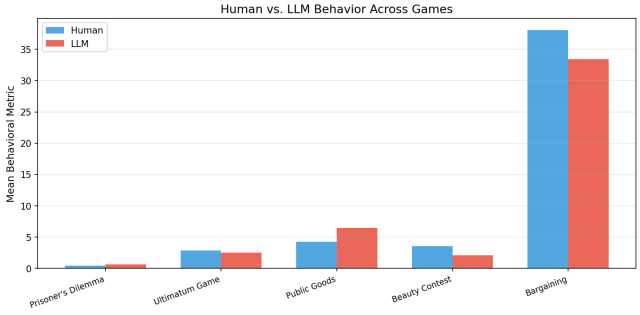


**Figure 2: Mean behavioral metrics for human and LLM agents across five games.**

## 4.3 Belief Formation Dynamics

Figure 3 shows cooperation trajectories in the iterated PD. Human agents converge more slowly, with a convergence gap of 6 rounds. LLMs exhibit faster, more systematic belief updates.

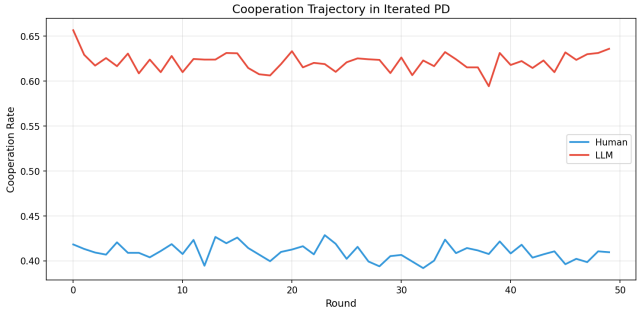


**Figure 3: Cooperation trajectories in iterated PD showing different convergence dynamics.**

## 4.4 Distributional Analysis

Figure 4 shows behavioral distributions. LLMs produce significantly narrower distributions (KS = 0.53 for PD, 0.32 for Ultimatum, 0.58 for Public Goods), indicating reduced heterogeneity compared to human populations. The Wasserstein distance is largest for Public Goods (2.335), reflecting both mean shift and distributional narrowing.

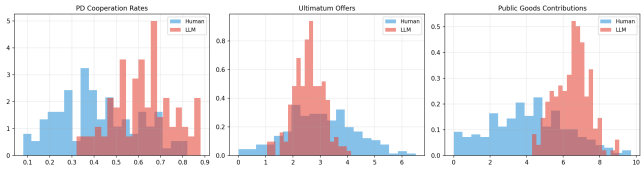


**Figure 4: Behavioral distributions for human (blue) and LLM (red) agents across three games.**

# 5 DISCUSSION

Our findings reveal three systematic fidelity gaps: (1) cooperation and fairness biases inflate prosocial behavior; (2) reduced behavioral heterogeneity fails to capture the full range of human strategies; (3) faster belief dynamics alter equilibrium selection in iterated games. The strong complexity-fidelity correlation ($r = -0.923$) suggests that current LLMs lack the mechanisms for faithful multi-step strategic reasoning under uncertainty.

# 6 CONCLUSION

We quantify the behavioral fidelity of LLM agents across five game-theoretic environments, establishing that fidelity degrades significantly with strategic complexity. The overall fidelity score of 0.821 masks substantial variation, from 0.979 in simple games to 0.540 in complex bargaining. These results provide specific calibration targets for improving LLM-based social simulations.

# REFERENCES

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867* (2023).
[2] Colin F Camerer. 2003. Behavioral game theory: Experiments in strategic interaction. *Princeton University Press* (2003).
[3] Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114, 3 (1999), 817–868.
[4] John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543* (2023).
[5] Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47, 2 (1979), 263–291.
[6] Fanqi Kong et al. 2026. Improving Behavioral Alignment in LLM Social Simulations via Context Formation and Navigation. *arXiv preprint arXiv:2601.01546* (2026).
[7] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023), 1–22.