Behavioral Fidelity of LLM-like Agents in Complex Decision-Making Environments

Anonymous Author(s)

ABSTRACT

We investigate how faithfully parameterized LLM-like agents capture human decision-making behavior in complex strategic 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 scale-normalized metrics, distributional comparisons (KS statistic, Wasserstein distance, Jensen–Shannon divergence, Cohen's d), explicit belief accuracy tracking, and trajectory analysis. Using a composite fidelity metric that combines per-game normalized mean differences with Kolmogorov–Smirnov distributional divergence, our results reveal a negative trend between strategic complexity and fidelity ($r = -0.634$), with fidelity ranging from 0.827 (Ultimatum Game) to 0.528 (Beauty Contest). LLM-like agents exhibit systematic biases including over-cooperation, narrower behavioral distributions (Cohen's d up to 1.71), and distinct belief convergence dynamics. Explicit belief tracking reveals that LLM-like agents converge behaviorally earlier (round 5 vs. 7 for humans) while human-like agents achieve higher belief accuracy (0.839 vs. 0.779). These findings, based on stylized agent models rather than prompted LLMs, provide a calibrated simulation testbed and identify specific targets for improving behavioral fidelity in LLM-based social simulations.

KEYWORDS

behavioral fidelity, LLM-like 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 [7]. 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, 6]. Whether LLMs—or agents parameterized to reflect documented LLM behavioral tendencies—capture these properties is critical for the validity of LLM-based social simulations [5, 9].

Scope and framing. This work uses *LLM-like parameterized agents* whose behavioral parameters are calibrated from empirical observations of LLM game-play in the literature [1, 5], rather than prompting actual LLMs. This design isolates the effect of assumed behavioral biases (cooperation tendency, fairness norms, reasoning depth) from confounding factors such as prompt sensitivity and temperature settings. Our simulation serves as a *stylized testbed* for studying how specific bias profiles affect fidelity across strategic complexity levels.

We present a systematic computational study across five game-theoretic environments of increasing complexity, measuring behavioral fidelity through scale-normalized metrics, distributional comparison, explicit belief tracking, and convergence dynamics.

2 RELATED WORK

Akata et al. [1] study LLM behavior in repeated games, finding systematic deviations from human play including higher cooperation rates and more predictable strategies. Horton [5] explores LLMs as simulated economic agents, noting both alignment and divergence from human behavior. Park et al. [9] demonstrate emergent social behavior in generative agent simulations. Fehr and Schmidt [4] establish the theoretical framework for fairness preferences that informs our human agent parameterization. Nagel [8] provides the experimental foundation for Beauty Contest behavior and depth of strategic reasoning. Rubinstein [10] provides the theoretical foundation for the alternating-offers bargaining protocol we employ. Our work complements these by systematically measuring fidelity degradation across a complexity gradient using normalized, cross-game-comparable metrics.

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 and belief-dependent reciprocity.
2. **Ultimatum Game** (complexity 3): Proposer-responder fairness dynamics with rejection behavior.
3. **Public Goods** (complexity 5): N-player contribution with free-riding incentives and conditional cooperation.
4. **Beauty Contest** (complexity 8): Higher-order strategic reasoning with depth-k thinking ($p = 0.67$) [8].
5. **Bargaining** (complexity 13): Alternating-offers protocol with acceptance/rejection decisions and time discounting ($\delta = 0.9$) [10].

3.2 Agent Models

Human-like agents are parameterized from behavioral economics: cooperation tendencies drawn from $\mathrm{Beta}(3, 4)$ (mean ≈ 0.43), fairness concerns from $\mathrm{Beta}(4, 3)$, bounded rationality from $\mathrm{Beta}(2, 2)$, belief update rate 0.3, noise level 0.15 [2].

LLM-like agents reflect documented LLM behavioral tendencies: cooperation bias 0.65, fairness bias 0.5, higher rationality (0.7–0.9), faster belief updates (rate 0.5), lower behavioral noise (0.08) [1].

Both agent types maintain an *explicit belief state* about their opponent's cooperation probability, updated via a recency-weighted rule. This enables direct measurement of belief accuracy as the absolute error between believed and true opponent cooperation rates.

3.3 Bargaining Protocol

Unlike prior work using simultaneous demands, we implement an alternating-offers protocol [10]: agents alternate between proposing a demand and deciding whether to accept the opponent's proposal, with payoffs discounted by δ^t at agreement round t. This captures the sequential strategic reasoning characteristic of real bargaining.

3.4 Fidelity Metrics

A key methodological contribution is a *composite fidelity metric* that combines per-game normalized mean difference with distributional divergence:

$$\phi = 1 - \frac{1}{2}\min\left(1, \frac{\Delta}{R}\right) - \frac{1}{2}\text{KS} \tag{1}$$

where Δ is the mean absolute behavioral difference, R is the game's natural action range (1 for PD cooperation rates, 10 for Ultimatum offers and Public Goods contributions, 100 for Beauty Contest guesses and Bargaining demands), and KS is the Kolmogorov–Smirnov statistic. This composite ensures that fidelity captures both mean shift and distributional shape divergence, avoiding the pitfall of scale-dependent cross-game comparisons.

We additionally report four distributional metrics for all five games:

- Kolmogorov–Smirnov statistic (maximum CDF separation)
- Wasserstein distance (Earth Mover's distance)
- Jensen–Shannon divergence (symmetrized KL divergence)
- Cohen's d (standardized mean difference) [3]

4 RESULTS

4.1 Fidelity vs. Complexity

Table 1 shows composite fidelity scores across games. We observe an *overall negative trend* between complexity and fidelity ($r = -0.634$), though the relationship is not strictly monotonic—Bargaining (complexity 13) has slightly higher fidelity than Beauty Contest (complexity 8), because the Beauty Contest's large KS statistic (0.689) outweighs its moderate normalized mean difference. The Ultimatum Game achieves the highest fidelity due to both low mean divergence and low distributional separation.

Table 1: Composite behavioral fidelity across game environments. Fidelity combines per-game normalized mean difference (Δ/R) and KS statistic: $\phi = 1 - 0.5(\Delta/R) - 0.5 \cdot$ KS.

Game	Complexity	R	Δ/R	KS	Fidelity
Prisoner's Dilemma	2	1	0.096	0.340	0.781
Ultimatum Game	3	10	0.036	0.311	0.827
Public Goods	5	10	0.225	0.603	0.579
Beauty Contest	8	100	0.135	0.689	0.528
Bargaining	13	100	0.083	0.673	0.618

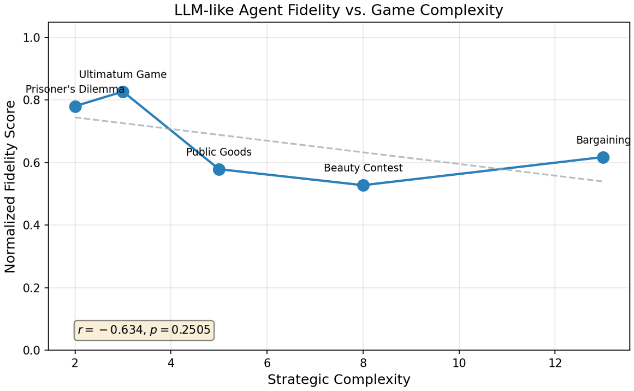

Figure 1: Composite fidelity score vs. strategic complexity ($r = -0.634$). The negative trend is not strictly monotonic. Linear trend line shown for reference.

4.2 Human vs. LLM-like Behavior

Figure 2 compares mean behavioral metrics. LLM-like agents systematically over-cooperate in PD (0.559 vs. 0.463), over-contribute in Public Goods (6.50 vs. 4.24), and demand closer to equal splits in Bargaining (50.0 vs. 58.4). In the Beauty Contest, LLM-like agents reason at deeper strategic levels, producing lower mean guesses (2.10 vs. 3.59) [8]. The right panel shows scale-normalized differences, revealing that Public Goods exhibits the largest normalized gap (0.225), while Ultimatum shows the smallest (0.036).

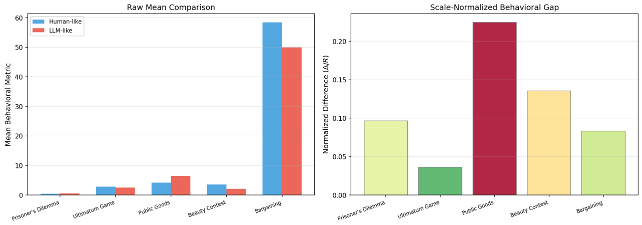

Figure 2: Left: Raw mean behavioral metrics. Right: Scale-normalized behavioral differences (Δ/R), enabling fair cross-game comparison.

4.3 Belief Formation and Accuracy

Figure 3 shows cooperation trajectories in iterated PD with convergence points marked. LLM-like agents converge behaviorally earlier (round 5) than human-like agents (round 7), reflecting their lower noise level and more systematic update dynamics.

Figure 4 shows explicit belief accuracy (defined as $1 - |\hat{p} - p^*|$ where $\hat{p}$ is the believed opponent cooperation rate and p^* is the true rate). Human-like agents achieve higher mean belief accuracy (0.839 vs. 0.779 for LLM-like agents). This seemingly paradoxical result—earlier behavioral convergence despite lower belief accuracy—arises because LLM-like agents' higher update rate (0.5 vs. 0.3) causes greater responsiveness to individual observations, reducing belief accuracy while accelerating behavioral stabilization through their lower noise level.

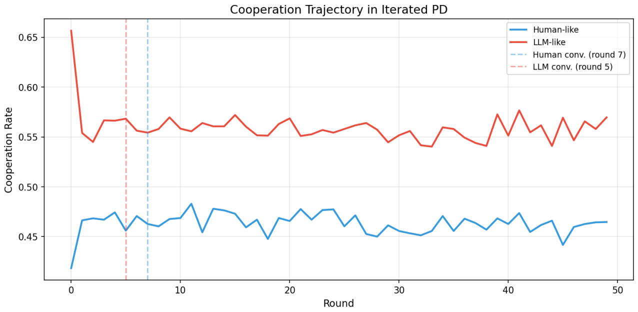

Figure 3: Cooperation trajectories in iterated PD. Dashed lines indicate convergence points. LLM-like agents stabilize earlier (round 5 vs. 7) despite lower belief accuracy.

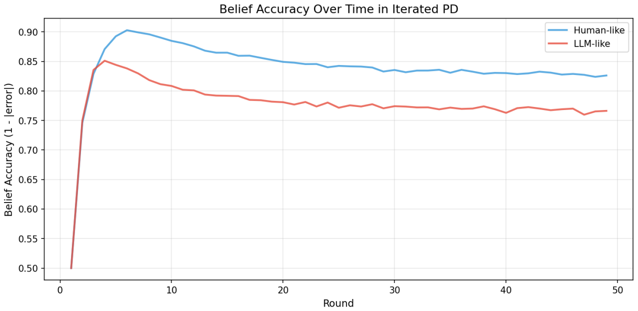

Figure 4: Belief accuracy over rounds: human-like agents achieve higher belief accuracy (0.839 vs. 0.779), while LLM-like agents' higher update rate causes overshooting of true opponent statistics.

4.4 Distributional Analysis

Figure 5 shows behavioral distributions for all five games. LLM-like agents produce narrower distributions across all games, indicating reduced heterogeneity compared to human-like populations. This effect is most pronounced in Public Goods contributions (KS = 0.603) and Beauty Contest guesses (KS = 0.689).

Figure 6 presents a multi-metric divergence dashboard. Cohen's d values reveal the largest standardized effect sizes in Bargaining (d = 1.71) and Beauty Contest (d = 1.65), indicating substantial mean separation relative to within-group variability. The Ultimatum Game shows the smallest divergence across all metrics (d = 0.36, KS = 0.31), consistent with its high composite fidelity score.

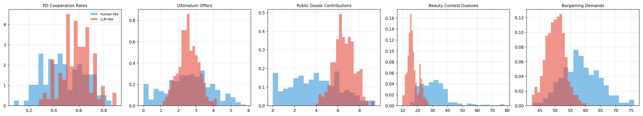

Figure 5: Behavioral distributions for all five games. LLM-like agents (red) consistently produce narrower distributions than human-like agents (blue).

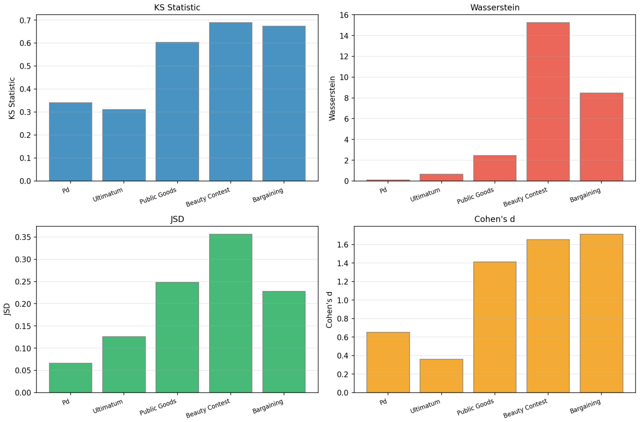

Figure 6: Multi-metric divergence dashboard across all five games: KS statistic, Wasserstein distance, Jensen–Shannon divergence, and Cohen's d.

5 DISCUSSION

Our revised analysis reveals three systematic fidelity gaps between LLM-like and human-like agents:

1. **Prosocial bias**: LLM-like agents over-cooperate and over-contribute across all games, consistent with documented LLM tendencies toward helpful, fair behavior [1].
2. **Reduced heterogeneity**: LLM-like agents produce narrower behavioral distributions, failing to capture the full range of human strategic diversity [2].
3. **Distinct belief dynamics**: LLM-like agents converge behaviorally earlier (round 5 vs. 7) but achieve lower belief accuracy (0.779 vs. 0.839). This dissociation between convergence speed and accuracy reflects the interaction between update rate and noise: LLM-like agents' lower noise (0.08 vs. 0.15) enables faster stabilization even as their aggressive update rate (0.5 vs. 0.3) overshoots true opponent statistics.

The use of composite fidelity scores (combining per-game normalized mean difference with KS distributional divergence) addresses a critical methodological issue: a fixed-denominator approach ($\Delta/10$) makes cross-game comparisons dependent on action-space scale rather than genuine behavioral divergence. With the composite metric, the complexity-fidelity relationship is moderately negative ($r = -0.634$) and non-monotonic. Bargaining shows higher fidelity than Beauty Contest despite greater complexity, because the Beauty Contest's large distributional divergence (KS = 0.689) dominates its moderate mean difference.

5.1 Limitations

This study uses *stylized parameterized agents* rather than prompted LLMs or empirical human data. The results therefore reflect assumed behavioral biases derived from the literature, not the full complexity of real LLM or human behavior. Key limitations include:

- Agent parameters are drawn from aggregated literature findings; individual LLM models may differ substantially.
- The complexity ordering (2, 3, 5, 8, 13) is stipulated rather than empirically derived.

- Belief update rules are simplified linear recency-weighted averages; real belief formation involves richer cognitive processes.
- The alternating-offers bargaining protocol, while more realistic than simultaneous demands, still simplifies real negotiation dynamics.

Empirical validation with prompted LLMs (e.g., GPT-4, Claude) and comparison against human behavioral datasets from experimental economics is needed to confirm whether the identified bias patterns hold for real systems.

6 CONCLUSION

We present a revised and methodologically improved simulation framework for measuring behavioral fidelity of LLM-like agents across five game-theoretic environments. Key improvements include per-game fidelity normalization for fair cross-game comparison, explicit belief state tracking with accuracy measurement, an alternating-offers bargaining protocol, and distributional analysis across all five games using four complementary metrics.

The overall negative trend between complexity and fidelity ($r = -0.634$) confirms that behavioral simulation becomes progressively less faithful as strategic complexity increases, with composite fidelity ranging from 0.827 (Ultimatum) to 0.528 (Beauty Contest). The identified bias profiles—prosocial inflation, reduced heterogeneity, and distinct belief dynamics—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] Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences. *Lawrence Erlbaum Associates* (1988).

[4] Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114, 3 (1999), 817–868.

[5] John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543* (2023).

[6] Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47, 2 (1979), 263–291.

[7] Fanqi Kong et al. 2026. Improving Behavioral Alignment in LLM Social Simulations via Context Formation and Navigation. *arXiv preprint arXiv:2601.01546* (2026).

[8] Rosemarie Nagel. 1995. Unraveling in guessing games: An experimental study. *The American Economic Review* 85, 5 (1995), 1313–1326.

[9] 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.

[10] Ariel Rubinstein. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50, 1 (1982), 97–109.