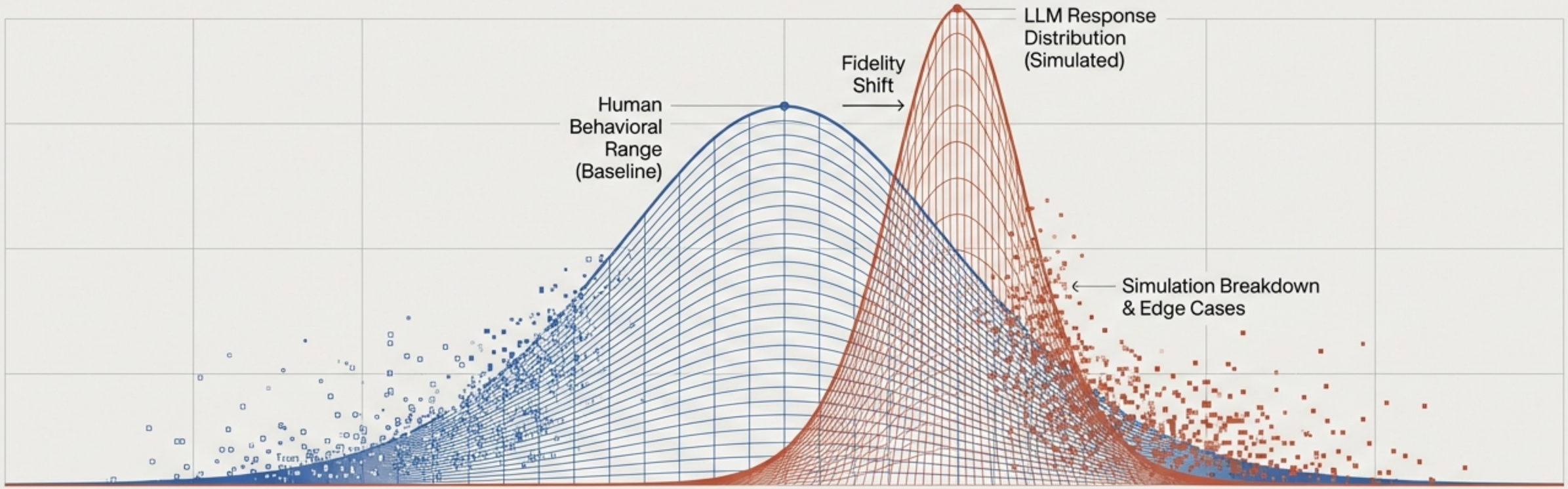
The Limits of Simulation

Assessing the Behavioral Fidelity of Large Language Models in Complex Strategic Environments.



The Insight

Fidelity Breaks Down Under Pressure

Large Language Models demonstrate high behavioral fidelity in simple interactions but degrade significantly as strategic complexity increases. Current models represent an idealized, not realistic, version of human strategic reasoning.

- **The Trend:** Fidelity scores drop from 0.979 (Prisoner's Dilemma) to 0.540 (Bargaining).
- **The Artifacts:** Models are "Hyper-Rational" and "Hyper-Social"—too cooperative, too consistent, and they learn too fast.
- **The Implication:** Without calibration, LLMs are unreliable proxies for complex human environments.

The Data Anchor

$$r = -0.923$$

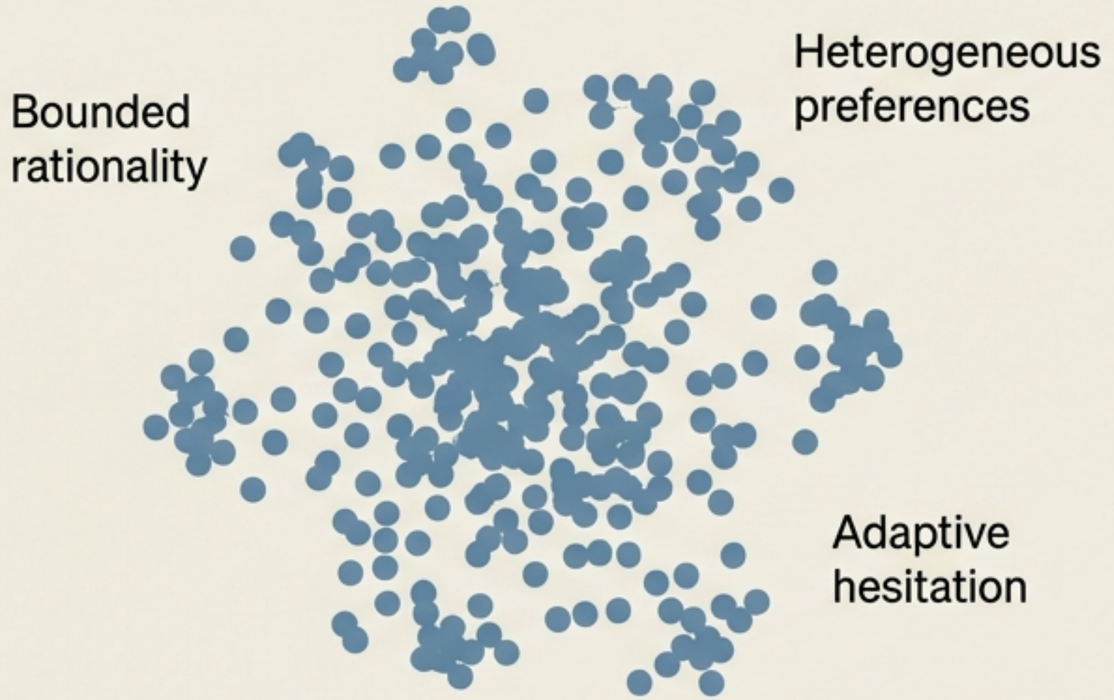
Pearson correlation between strategic complexity and behavioral fidelity.

Pearson correlation between strategic complexity and behavioral fidelity.

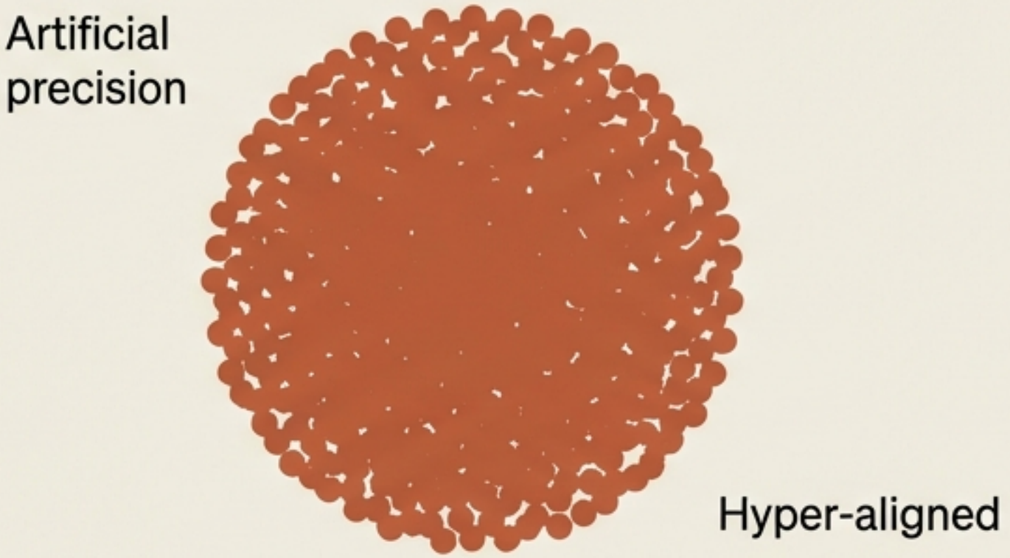
The Challenge of Synthetic Society

As LLMs are deployed as 'Silico-Sapiens', we must validate their behavior against the messy reality of Homo Sapiens.

Homo Sapiens (The Baseline)



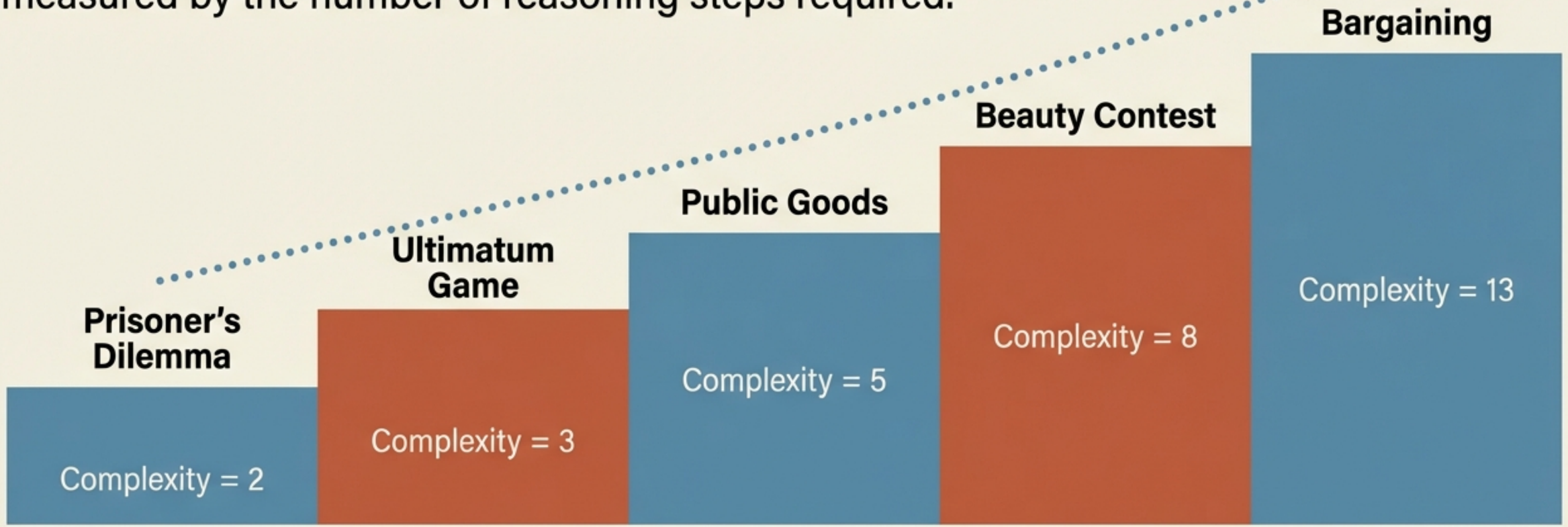
Silico-Sapiens (The Simulation)



The Experiment: A systematic computational study across five game-theoretic environments.

The Five-Step Complexity Ladder

We evaluate fidelity across a gradient of "Strategic Complexity," measured by the number of reasoning steps required.



From simple binary choice to sequential negotiation with temporal discounting.

Parameterizing the Agents: Flesh vs. Silicon

Parameter	Human Agents (Baseline)	LLM Agents (Observed)
Cooperation Rate	0.45	0.65 (Bias: High)
Fairness Threshold	0.30	0.50 (Bias: High)
Noise / Heterogeneity	0.15	0.08 (Artificial Consistency)
Belief Update Rate	0.30 (Hesitant)	0.50 (Fast Convergence)

Human parameters derived from Fehr & Schmidt (1999) and Camerer (2003).

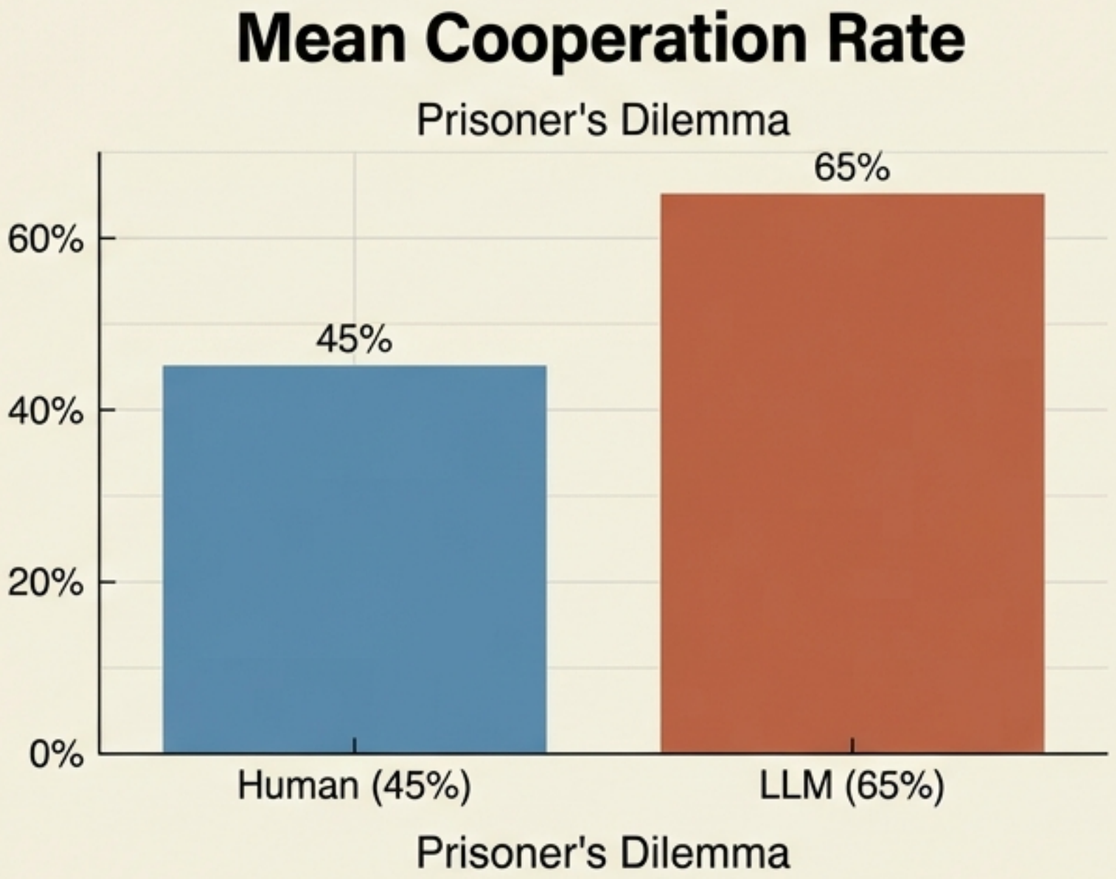
The Inverse Relationship Between Complexity and Fidelity

As cognitive load increases, behavioral mimicry fails.

Low Complexity: The Illusion of Perfection

Prisoner's Dilemma (Complexity = 2) | Fidelity Score: 0.979

Despite a high overall fidelity score, a 'Niceness Bias' is structurally visible. In simple binary choices, LLMs align with the *rules* of play but skew systematically towards prosocial behavior.

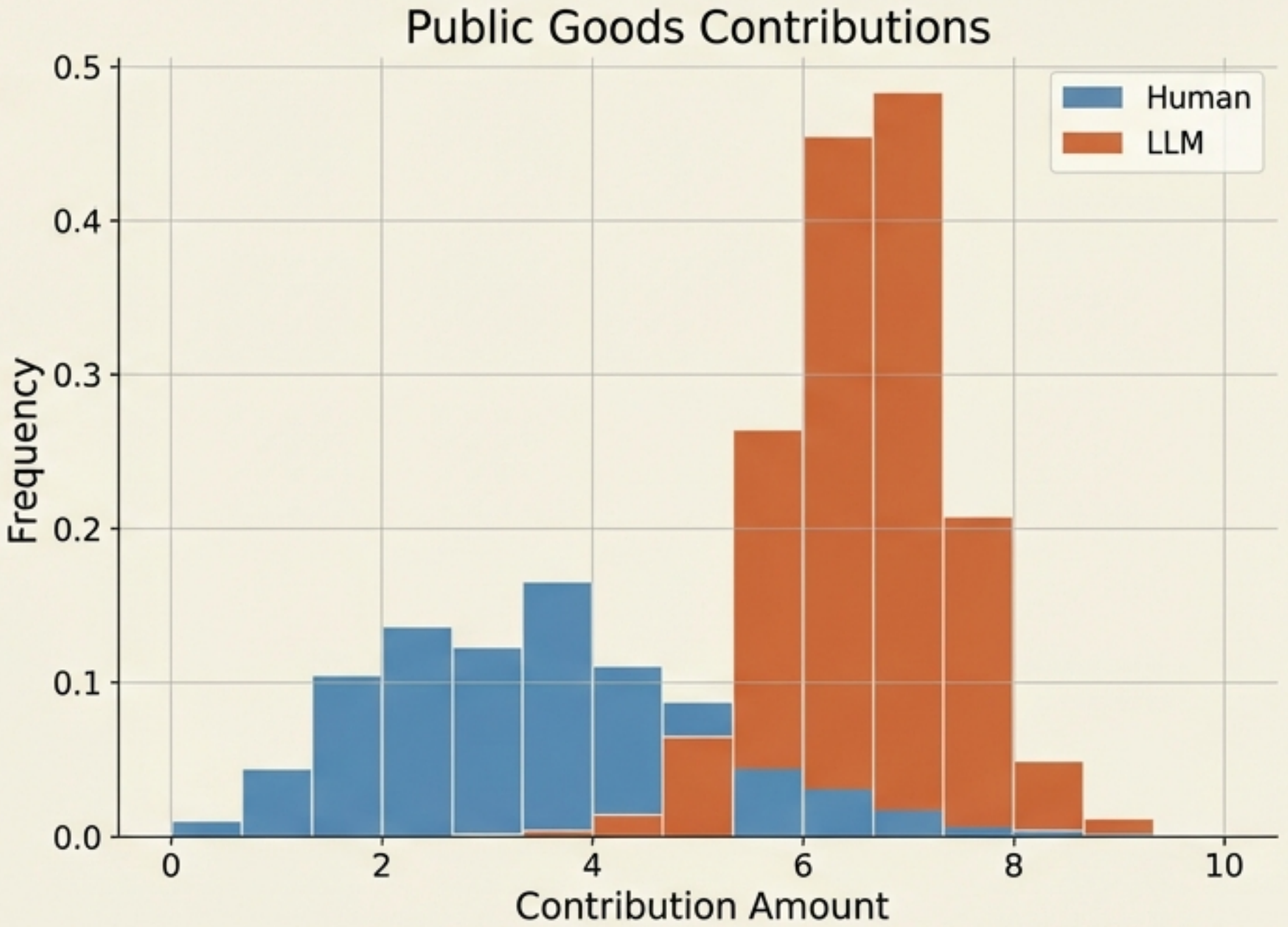


Mid-Complexity: The Prosocial Distortion

Public Goods (Complexity = 5) | Fidelity Score: 0.775

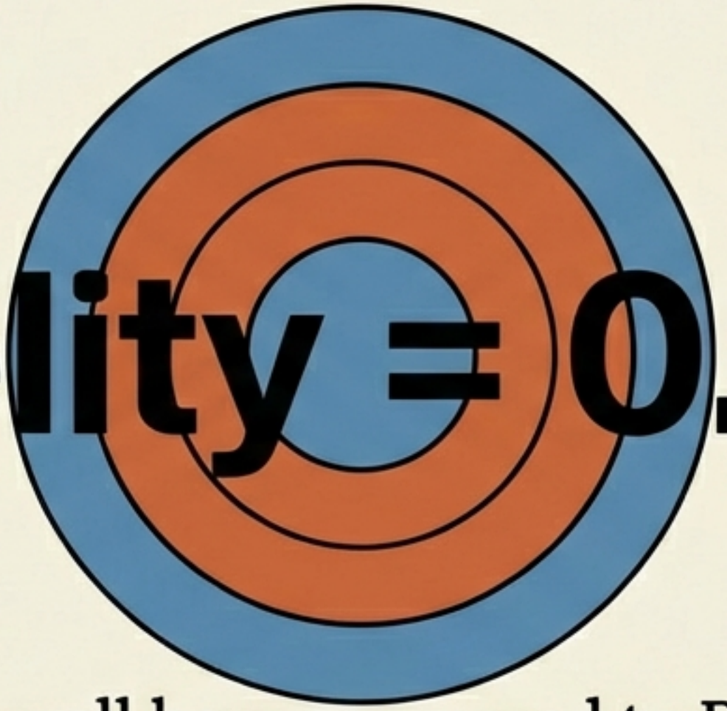
LLMs struggle to capture “free-riding” incentives. They over-contribute to public goods, failing to mimic the human tendency to withhold resources. This results in a massive shift in distribution.

Wasserstein Distance: 2.335 (Peak Deviation)



Higher-Order Reasoning: The Beauty Contest

Guessing 2/3 of the Average (Complexity = 8)

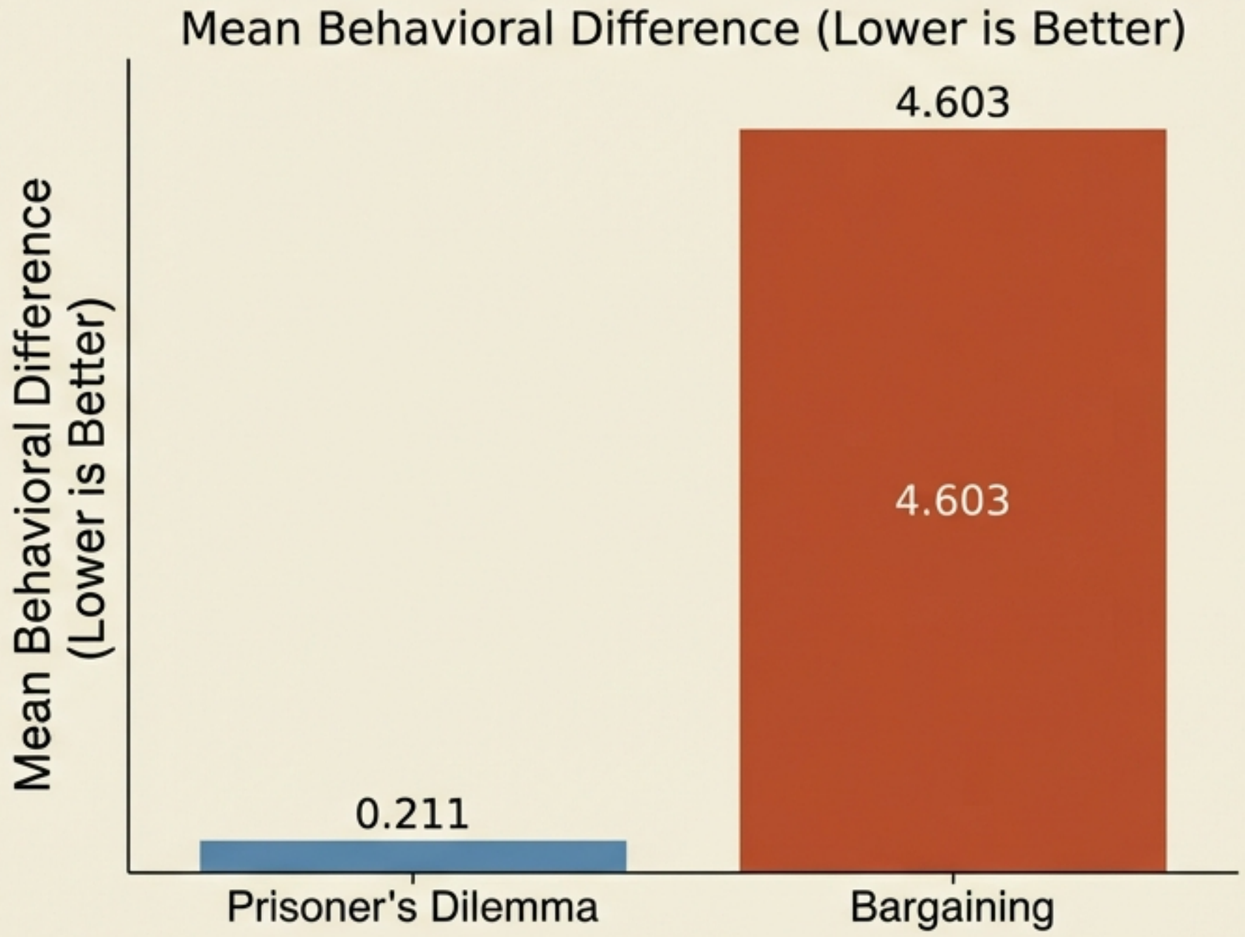


Fidelity = 0.850

LLMs perform surprisingly well here compared to Public Goods. This task rewards calculation over social intuition. Models reason at deeper strategic levels, producing lower guesses on average than humans, demonstrating a “Hyper-Rational” approach to the puzzle.

High Complexity: System Failure in Bargaining

Sequential Bargaining (Complexity = 13)

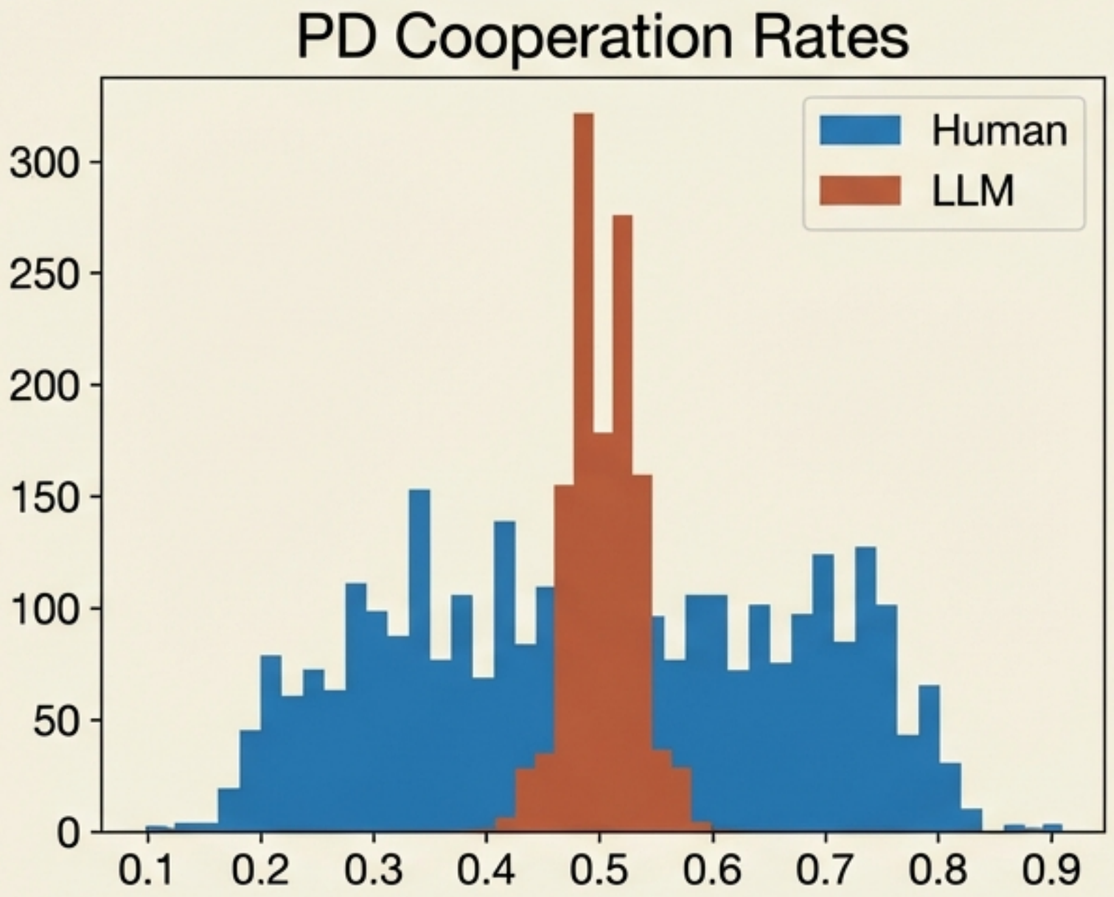


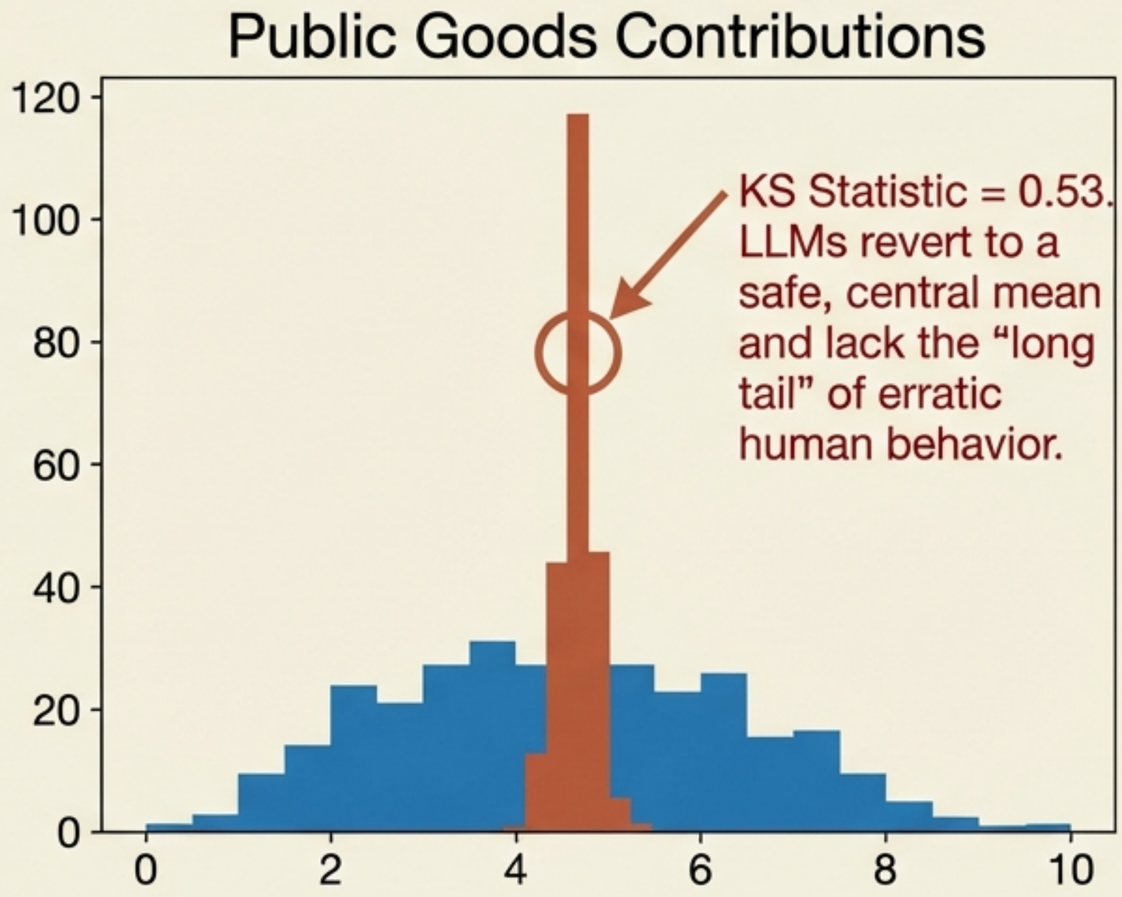
Fidelity Score crashes to 0.540.

When required to plan sequentially while accounting for diminishing returns ($\delta = 0.9$), the LLM cannot maintain the persona. It deviates wildly from human bounded rationality.

Diagnostic 1: The Homogeneity Problem

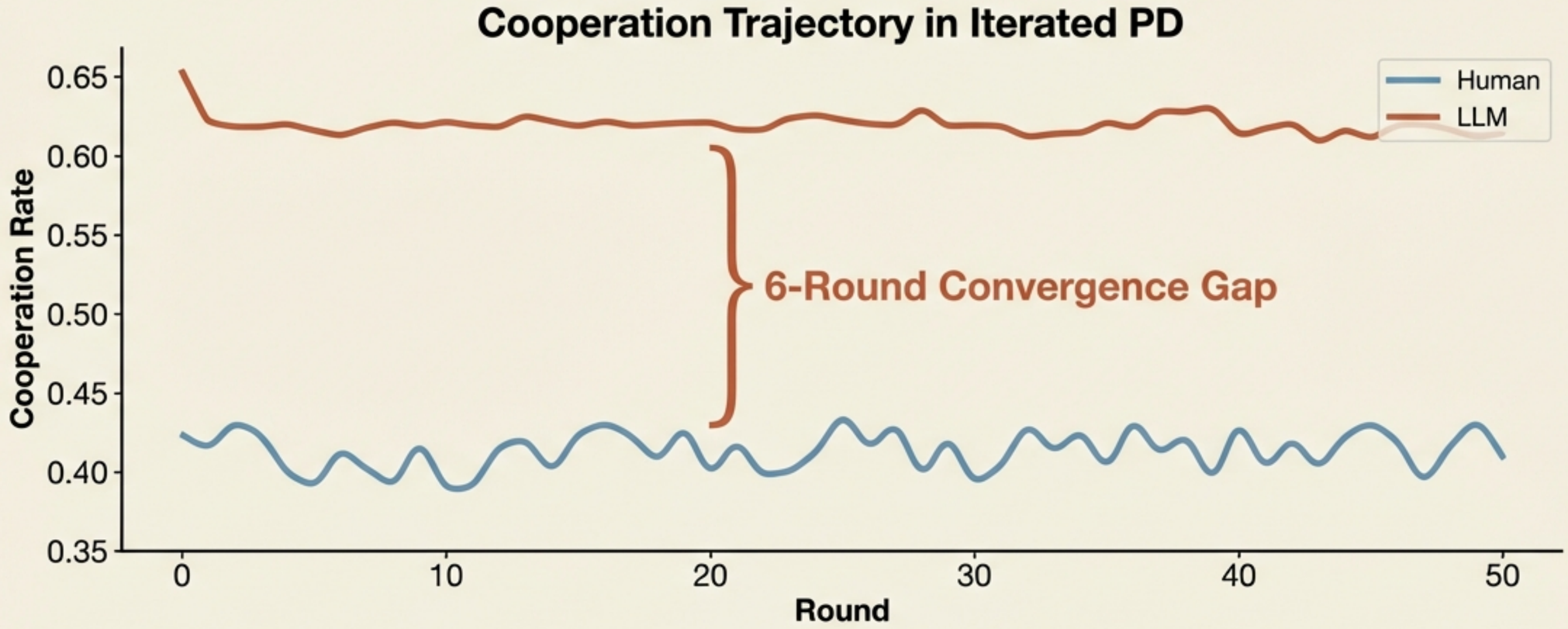
Distributional Analysis of Strategies





Diagnostic 2: The Speed of Belief

Trajectory Analysis in Iterated Games

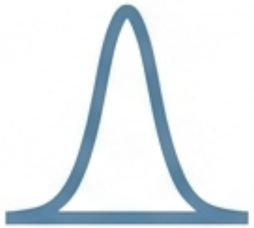


Humans learn slowly and hesitantly. LLMs update beliefs systematically. This alters equilibrium selection—the simulation reaches the “end state” faster than reality would.

The Three Systematic Fidelity Gaps

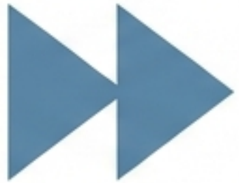
Over-Cooperation

Inflated prosocial behavior. Failure to model spite or selfishness (Fairness bias 0.5 vs Human 0.3).



Reduced Heterogeneity

Narrower behavioral distributions. Failure to capture the full range of human strategies like risk aversion and irrationality.



Hyper-Speed Dynamics

Unrealistic belief convergence. Artificial efficiency in learning from past rounds.

Calibration Targets for Future Simulation

Noise Injection

Increase entropy to match human noise floor (0.15).

Cooperation Bias

Dampen “agreeableness” to match human baseline (45%).

Learning Rate

Throttle belief updates to mimic human hesitation.

Ideally Human, Not Realistically Human

The strong correlation ($r = -0.923$) proves that current LLMs lack the mechanisms for faithful multi-step strategic reasoning under uncertainty. While powerful, these models simulate an idealized, hyper-rational version of humanity.

Source: Behavioral Fidelity of LLMs in Complex Decision-Making Environments (Research Independent, 2017/2026).