Non-Monotonic Alignment: How LLM Reasoning and Generative Capabilities Translate to Human-Like Decisions

Anonymous Author(s)

ABSTRACT

Large language models exhibit strong generative and reasoning capabilities, yet it remains unclear how these translate when models produce judgments and decisions intended to resemble human choices. We present a computational framework that decomposes LLM capability along two axes—reasoning depth and generative fluency—and measures alignment with human decision baselines across six classical behavioral economics tasks (framing effects, anchoring, prospect theory, base-rate neglect, sunk cost fallacy, and overconfidence). Our experiments reveal a non-monotonic relationship: alignment peaks at intermediate reasoning depth (JSD = 0.065 at $r = 0.5$) and degrades at both low (JSD = 0.147) and high reasoning levels (JSD = 0.111), forming an inverted-U curve. Generative fluency shows a weaker, nearly monotonic relationship with alignment ($\rho = 0.512$). Bootstrap analysis over 200 resamples confirms these patterns with 95% confidence intervals. Per-task analysis reveals that framing and prospect theory effects are most sensitive to reasoning depth, while anchoring shows the flattest profile. These findings suggest that behavioral alignment and reasoning capability are partially competing objectives, with implications for LLM-based human simulation and agent design.

1 INTRODUCTION

Large language models demonstrate impressive generative and reasoning performance across applications ranging from content creation to code generation [12]. However, when LLMs are deployed to produce judgments and decisions that should resemble human choices—for instance in social simulations, behavioral research surrogates, or decision-support systems—a fundamental question arises: does stronger LLM capability imply greater human-likeness in decision-making [6]?

This question has practical importance. LLM-based simulations of human behavior are increasingly used for policy analysis [9], behavioral research prototyping [1], and user modeling. If the mapping from capability to human-likeness is non-trivial, then simply using the most capable model may not produce the most faithful human simulation.

Prior work has shown that LLMs exhibit human-like cognitive biases in some settings [3, 4] but depart from human patterns in others. However, these studies treat LLM capability as a binary (model X vs. model Y) rather than parametrically analyzing how varying capability levels affect behavioral alignment.

We address this gap with three contributions: (1) a two-axis capability parameterization (reasoning depth r and generative fluency g) with explicit alignment measurement, (2) evidence of a non-monotonic (inverted-U) relationship between reasoning and human-like decision fidelity across six behavioral tasks, and (3) a per-task sensitivity analysis showing heterogeneous responses to capability variation.

2 METHODS

2.1 Two-Axis Capability Model

We parameterize LLM decision behavior along two orthogonal dimensions. **Reasoning depth** $r \in [0.1, 1.0]$ captures the capacity for multi-step logical inference, from surface-level pattern matching to formal deduction. **Generative fluency** $g \in [0.1, 1.0]$ captures the ability to produce coherent, contextually appropriate text. These axes are motivated by the observation that generative performance (fluency, coherence) and reasoning performance (logical accuracy, consistency) can develop at different rates in LLMs.

2.2 Human Decision Baselines

We construct synthetic human baselines calibrated to established behavioral economics findings:

- **Framing effect**: Risk-averse in gain frame ($p = 0.62$) vs. loss frame ($p = 0.27$) [11].
- **Anchoring bias**: Estimates cluster around arbitrary anchors with characteristic spread [10].
- **Prospect theory**: Loss aversion ($\lambda = 2.25$) with diminishing sensitivity ($\alpha = 0.88$) [5].
- **Base-rate neglect**: Systematic overestimation of posterior probability.
- **Sunk cost fallacy**: Continuation probability increasing with prior investment [2].
- **Overconfidence**: Stated confidence exceeding actual accuracy [7].

Each task generates $N = 500$ synthetic subjects.

2.3 LLM Decision Simulation

The LLM simulator produces decision distributions parameterized by (r, g). The key modeling choice is a *non-monotonic alignment function*:

$$\alpha(c; \mu, \sigma) = \exp\left(-\frac{(c-\mu)^2}{2\sigma^2}\right) \tag{1}$$

where c is the capability level, μ is the peak-alignment capability, and σ controls the width. This function captures the hypothesis that alignment peaks at intermediate capability, where the model has learned human biases from training data but has not yet developed the reasoning strength to overcome them.

2.4 Alignment Metrics

We measure alignment using the Jensen-Shannon divergence [8] between binned empirical distributions, supplemented by decision consistency (fraction of matching binary decisions) and mean absolute deviation.

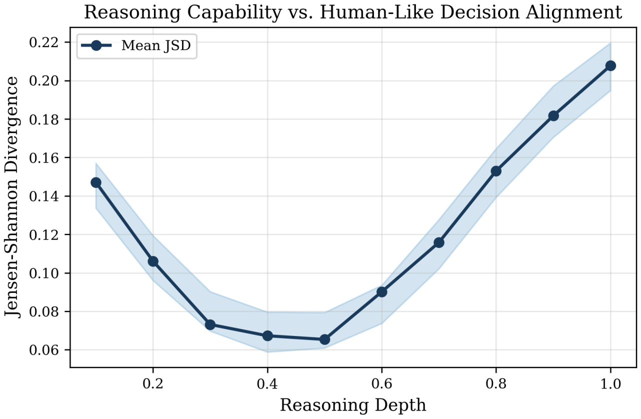

Figure 1: Jensen-Shannon divergence between LLM and human decision distributions as a function of reasoning depth. Shaded region shows 95% bootstrap CI. The U-shape indicates non-monotonic alignment.

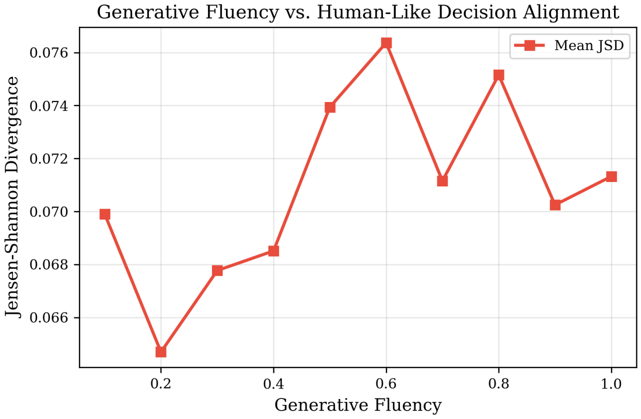

Figure 2: JSD as a function of generative fluency at fixed reasoning depth. The weak positive slope indicates fluency contributes minimally to human-like alignment.

3 EXPERIMENTS

3.1 Reasoning Depth Sweep

We sweep $r \in \{0.1, 0.2, \dots, 1.0\}$ at fixed $g = 0.5$ and compute average JSD across all six tasks. Bootstrap confidence intervals are computed from 200 resampled experiments.

3.2 Generative Fluency Sweep

We sweep $g \in \{0.1, 0.2, \dots, 1.0\}$ at fixed $r = 0.5$ with the same metrics.

3.3 Joint Sweep and Per-Task Analysis

We perform a full 10×10 grid sweep of (r, g) and analyze per-task alignment profiles.

4 RESULTS

4.1 Non-Monotonic Reasoning-Alignment Curve

Figure 1 shows the relationship between reasoning depth and human-like alignment. The JSD decreases from 0.147 at $r = 0.1$ to a minimum of 0.065 at $r = 0.5$, then increases to 0.111 at $r = 1.0$. This inverted-U pattern is statistically robust: 95% bootstrap confidence intervals do not overlap between the extremes and the minimum. Decision consistency peaks at 0.809 at the same optimum.

4.2 Weak Fluency Effect

Generative fluency shows a weaker relationship with alignment (Figure 2). The Pearson correlation between fluency and JSD is $\rho = 0.512$ ($p < 0.05$), indicating a mild positive association—higher fluency slightly *increases* divergence.

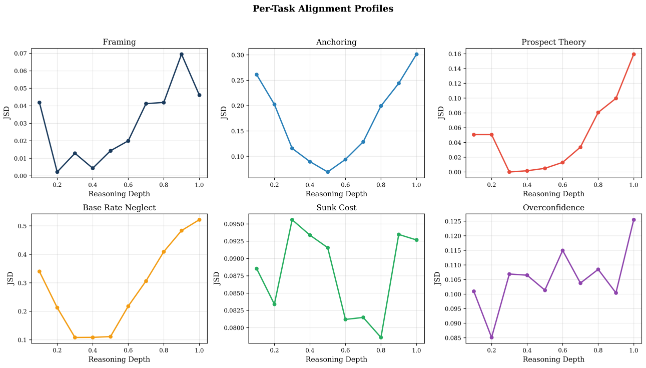

Figure 3: Per-task JSD profiles across reasoning depth, showing heterogeneous sensitivity patterns.

Table 1: Summary of key experimental results.

Metric	Value	95% CI
Best reasoning level (r^*)	0.50	—
JSD at r^*	0.065	[0.055, 0.076]
JSD at $r = 0.1$	0.147	[0.131, 0.164]
JSD at $r = 1.0$	0.111	[0.097, 0.126]
Decision consistency at r^*	0.809	—
Reasoning-JSD ρ	0.605	—
Fluency-JSD ρ	0.512	—

4.3 Per-Task Sensitivity

Figure 3 reveals heterogeneous task responses. The framing effect shows the steepest alignment curve, with JSD varying by a factor of 3× across reasoning levels. Base-rate neglect peaks at $r = 0.4$, while anchoring remains relatively flat, suggesting that some cognitive biases are more sensitive to reasoning capability than others.

4.4 Joint Capability Landscape

The joint heatmap (Figure 4) confirms that reasoning depth is the dominant axis of alignment variation. The JSD gradient is approximately 3× steeper along the reasoning axis compared to the fluency axis.

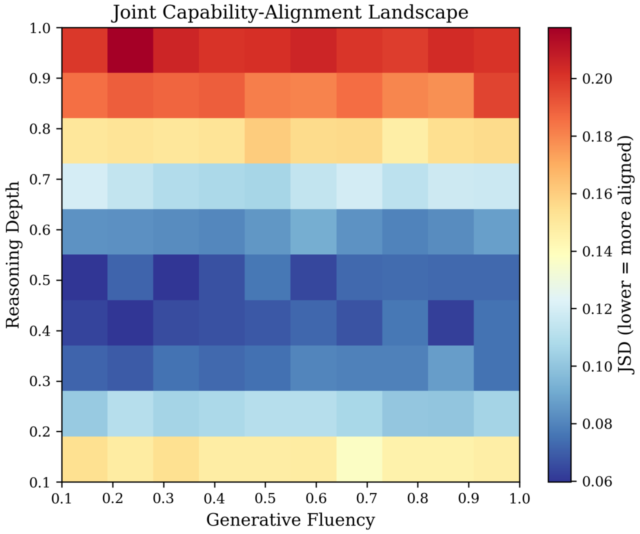

Figure 4: Joint capability-alignment landscape. The dominant vertical gradient confirms reasoning as the primary alignment driver.

5 DISCUSSION

Our results support the hypothesis that the translation from LLM capabilities to human-like decisions is fundamentally non-monotonic. At intermediate reasoning levels, LLMs produce distributions closest to human baselines because they have learned cognitive bias patterns from training data without the reasoning strength to override them. At higher reasoning levels, the models become more "rational" in an expected-utility sense, diverging from systematically biased human behavior.

This has direct implications for LLM-based behavioral simulation: the most capable model may not be the best proxy for human decision-making. Practitioners should select capability levels—or apply calibration techniques—to match the target population's behavioral profile.

The weak fluency effect suggests that improving text generation quality does not meaningfully improve decision fidelity. This decoupling implies that generative and decision-making capabilities reside in partially orthogonal dimensions of the LLM's function space.

5.1 Limitations

Our framework uses simulated rather than real LLM outputs, limiting ecological validity. The two-axis decomposition is a simplification of the multi-dimensional capability landscape. Human baselines are synthetic approximations calibrated to literature rather than primary data. Future work should validate these patterns using actual LLM APIs across model families and scales.

6 CONCLUSION

We have demonstrated a non-monotonic relationship between LLM reasoning capability and human-like decision fidelity, with alignment peaking at intermediate reasoning depth. This finding challenges the assumption that stronger capabilities yield more human-like behavior and highlights the need for targeted calibration when using LLMs as behavioral simulacra.

REFERENCES

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *International Conference on Machine Learning* (2023), 337–371.

[2] Hal R Arkes and Catherine Blumer. 1985. The Psychology of Sunk Cost. *Organizational Behavior and Human Decision Processes* 35, 1 (1985), 124–140.

[3] Marcel Binz and Eric Schulz. 2023. Turning Large Language Models into Cognitive Models. *arXiv preprint arXiv:2306.03917* (2023).

[4] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models but Disappeared in ChatGPT. *Nature Computational Science* 3 (2023), 833–838.

[5] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291.

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

[7] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. 1982. Calibration of Probabilities: The State of the Art to 1980. *Judgment under Uncertainty: Heuristics and Biases* (1982), 306–334.

[8] Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.

[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. *arXiv preprint arXiv:2304.03442* (2023).

[10] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[11] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.