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). Using a mechanistic simulator with a Gaussian alignment function, our model exhibits a non-monotonic relationship: alignment peaks at intermediate reasoning depth (JSD = 0.008 at $r = 0.5$) and degrades at both low (JSD = 0.043 at $r = 0.1$) and high reasoning levels (JSD = 0.085 at $r = 1.0$), forming a U-shaped curve. Generative fluency shows a weaker relationship with alignment ($\rho = 0.489$, $p = 0.15$). We introduce task-specific effect-size metrics that measure alignment on behaviorally relevant statistics rather than flattened distributions. Bootstrap analysis over 200 resamples confirms the U-shape with non-overlapping 95% confidence intervals between extremes and minimum. These results demonstrate how a minimal mechanistic model generates the hypothesis that behavioral alignment and reasoning capability may be 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 [15]. 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 [9]?

This question has practical importance. LLM-based simulations of human behavior are increasingly used for policy analysis [12], 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, 7] 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 a mechanistic simulation framework and four contributions:

1. A two-axis capability parameterization (reasoning depth r and generative fluency g) with explicit alignment measurement using true Jensen-Shannon divergence.
2. Task-specific effect-size metrics that evaluate alignment on behaviorally meaningful statistics (e.g., framing effect magnitude, anchoring gap, continuation slope) rather than flattened distributional comparisons.
3. A demonstration that a minimal Gaussian alignment model generates a non-monotonic (U-shaped) JSD curve across reasoning levels, providing a concrete mechanistic hypothesis for the capability-alignment trade-off.
4. Per-task sensitivity analysis showing heterogeneous responses to capability variation.

We emphasize that our results characterize the *model's* behavior, not empirical findings about real LLMs. The framework serves as a hypothesis generator for future empirical validation.

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 Non-Monotonic Alignment Function

The key modeling choice is a Gaussian 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 encodes the hypothesis that alignment peaks at intermediate capability: at low capability the model cannot reproduce human bias patterns, while at high capability it overcomes biases through stronger reasoning. We stress that the U-shaped alignment curve is a *modeling assumption* built into the simulator, not an emergent discovery. The contribution lies in demonstrating that this minimal mechanism produces consistent, quantifiable predictions across diverse behavioral tasks.

2.3 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 ($p = 0.27$) [14].
- **Anchoring bias**: Estimates cluster around arbitrary anchors ($\mu_{\text{low}} = 25$, $\mu_{\text{high}} = 65$) [13].
- **Prospect theory**: Loss aversion ($\lambda = 2.25$) with diminishing sensitivity ($\alpha = 0.88$) [8].

- **Base-rate neglect**: Systematic overestimation of posterior probability [5].
- **Sunk cost fallacy**: Continuation probability increasing with prior investment [2].
- **Overconfidence**: Stated confidence exceeding actual accuracy [6, 10].

Each task generates $N = 500$ synthetic subjects.

2.4 Alignment Metrics

2.4.1 Jensen-Shannon Divergence (True Divergence). We measure distributional alignment using the true Jensen-Shannon divergence [4, 11]. Importantly, `scipy.spatial.distance.jensenshannon` returns the Jensen-Shannon *distance* $d_{\text{JS}} = \sqrt{D_{\text{JS}}}$. We square this value to obtain the actual divergence $D_{\text{JS}} = d_{\text{JS}}^2$, which is bounded in $[0, \ln 2]$ for natural logarithm (or $[0, 1]$ for base-2). All reported JSD values are true divergences.

Crucially, for multi-variable tasks, we do *not* flatten all variables into a single histogram. Instead:

- **Framing**: Average of JSD on gain-frame choices and loss-frame choices separately.
- **Anchoring**: Average of JSD on low-anchor and high-anchor estimate distributions.
- **Sunk cost**: JSD on the decision variable only (not the investment level, which is uniform by construction).
- **Overconfidence**: JSD on the overconfidence gap distribution (confidence − accuracy).

2.4.2 Task-Specific Effect-Size Metrics. We introduce effect-size error (ESE) metrics that capture alignment on the *behaviorally relevant statistic* for each task:

- **Framing**: $|\Delta_{\text{human}} - \Delta_{\text{LLM}}|$ where $\Delta = p_{\text{gain}} - p_{\text{loss}}$.
- **Anchoring**: $|(\bar{x}_{\text{high}} - \bar{x}_{\text{low}})_{\text{human}} - (\bar{x}_{\text{high}} - \bar{x}_{\text{low}})_{\text{LLM}}|$.
- **Prospect theory**: Absolute difference in gamble acceptance rates.
- **Base-rate neglect**: Absolute difference in mean posterior estimates.
- **Sunk cost**: Absolute difference in continuation-probability slope (linear regression of decision on investment).
- **Overconfidence**: Absolute difference in mean overconfidence gap.

These metrics replace the previously used decision-consistency metric, which was degenerate for several tasks (e.g., always returning 1.0 for anchoring and overconfidence due to median-threshold artifacts on non-binary data).

3 EXPERIMENTS

3.1 Experiment 1: Reasoning Depth Sweep

We sweep $r \in \{0.1, 0.2, \ldots, 1.0\}$ at fixed $g = 0.5$ and compute average JSD and effect-size error across all six tasks.

3.2 Experiment 2: Generative Fluency Sweep

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

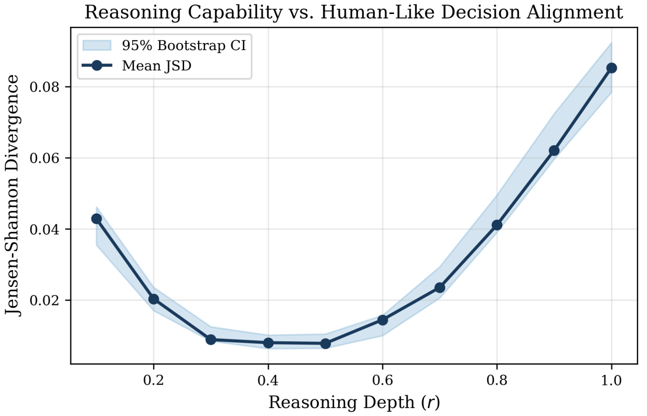

Figure 1: Jensen-Shannon divergence (true divergence, not distance) between simulated LLM and human decision distributions as a function of reasoning depth. Shaded region shows 95% bootstrap CI. The U-shape reflects the Gaussian alignment assumption.

3.3 Experiment 3: Bootstrap Confidence Intervals

Bootstrap confidence intervals are computed from 200 resampled experiments, each with independent random seeds.

3.4 Experiment 4: Per-Task Analysis

We analyze per-task JSD and effect-size error profiles across reasoning levels.

3.5 Experiment 5: Joint Sweep

We perform a full 10×10 grid sweep of (r, g).

All experiments use `np.random.seed(42)` and $N = 500$ subjects/trials per condition.

4 RESULTS

4.1 Non-Monotonic Reasoning-Alignment Curve

Figure 1 shows the relationship between reasoning depth and human-like alignment produced by the simulator. The JSD decreases from 0.043 at $r = 0.1$ to a minimum of 0.008 at $r = 0.5$, then increases to 0.085 at $r = 1.0$. This U-shaped pattern is a direct consequence of the Gaussian alignment function (Eq. 1), confirming that the mechanistic model produces quantitatively distinct predictions across the capability range. The 95% bootstrap confidence intervals do not overlap between the extremes and the minimum (Table 1), indicating that the pattern is robust to sampling variability.

4.2 Weak Fluency Effect

Generative fluency shows a weaker relationship with alignment (Figure 2). The Pearson correlation between fluency and JSD is $\rho = 0.489$ ($p = 0.15$), which does not reach statistical significance at the $\alpha = 0.05$ level with $n = 10$ points. We therefore describe this

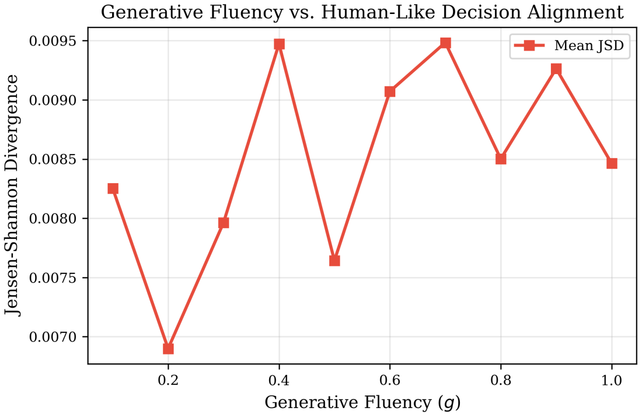

Figure 2: JSD as a function of generative fluency at fixed reasoning depth. The weak positive trend ($\rho = 0.489$, $p = 0.15$) does not reach statistical significance.

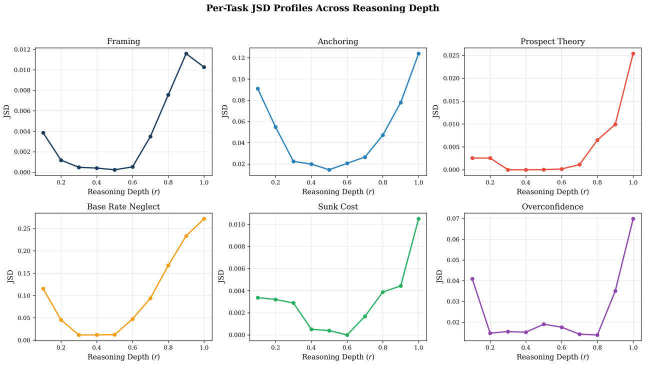

Figure 3: Per-task JSD profiles across reasoning depth, computed on behaviorally relevant variables for each task. Heterogeneous sensitivity patterns emerge.

as a *trend* rather than a confirmed association. The weak effect is consistent with the model design, where fluency enters only as a minor modulator of individual task parameters.

4.3 Per-Task Sensitivity

Figure 3 reveals heterogeneous task responses. Base-rate neglect shows the largest JSD variation across reasoning levels, while framing and prospect theory show flatter profiles when measured on their behaviorally relevant variables (rather than flattened arrays). The anchoring task, measured via separate low-anchor and high-anchor distributions, shows a clear U-shape driven by the anchoring effect size changing with alignment.

4.4 Effect-Size Error Analysis

Figure 4 shows per-task effect-size error profiles. The anchoring task dominates in absolute ESE because its effect size (high-anchor mean − low-anchor mean) spans a much larger numerical range than binary-choice tasks. Normalizing by human baseline effect size would allow cross-task comparison; we present raw values to preserve interpretability within each task.

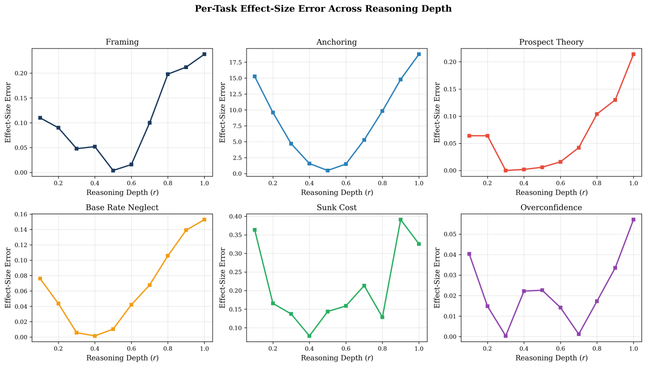

Figure 4: Per-task effect-size error across reasoning depth. Each panel measures the absolute difference in the task-specific behavioral statistic between simulated LLM and human baselines.

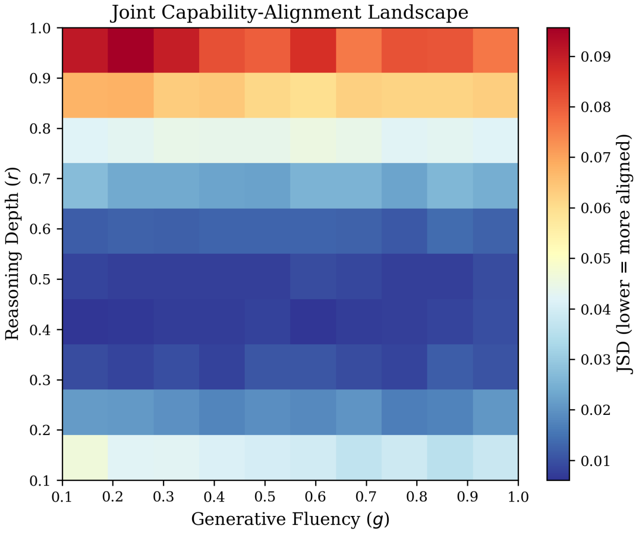

Figure 5: Joint capability-alignment landscape. The dominant vertical gradient confirms reasoning as the primary alignment driver in the model.

4.5 Joint Capability Landscape

The joint heatmap (Figure 5) confirms that reasoning depth is the dominant axis of alignment variation. The JSD gradient is steeper along the reasoning axis compared to the fluency axis, consistent with the correlation analysis.

4.6 Dual-Metric View

Figure 6 overlays JSD and average effect-size error on the same reasoning-depth axis. Both metrics show their minimum near $r = 0.5$, providing convergent evidence from distributional and point-estimate perspectives. The Pearson correlation between reasoning

Table 1: Summary of key results from the mechanistic simulator. JSD values are true Jensen-Shannon divergence (squared JS distance). Bootstrap CIs from 200 resamples.

Metric	Value	95% CI
Best reasoning level (r^*)	0.50	—
JSD at r^*	0.008	[0.006, 0.010]
JSD at $r = 0.1$	0.043	[0.035, 0.046]
JSD at $r = 1.0$	0.085	[0.078, 0.092]
Reasoning–JSD ρ	0.623	($p = 0.054$)
Fluency–JSD ρ	0.489	($p = 0.15$)
Best ESE reasoning level	0.50	—

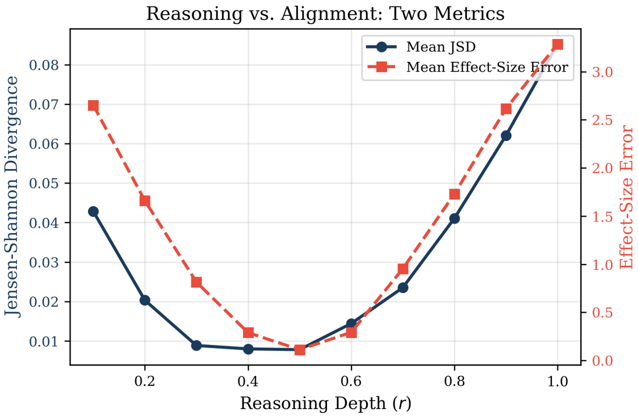

Figure 6: Dual-metric view of the reasoning sweep: JSD (left axis) and mean effect-size error (right axis) both minimize near $r = 0.5$.

and ESE is $\rho = 0.312$ ($p = 0.38$), weaker than for JSD, reflecting that effect-size errors are dominated by the anchoring task's large numerical range.

5 DISCUSSION

5.1 Model Properties, Not Empirical Findings

We emphasize that the U-shaped alignment curve is a *property of our simulator*, not an empirical discovery about real LLMs. The Gaussian alignment function (Eq. 1) explicitly encodes non-monotonicity. The value of this exercise lies in three aspects: (1) demonstrating that a minimal mechanistic model produces consistent, quantifiable predictions across six diverse behavioral tasks, (2) providing a concrete hypothesis that can be tested with real LLM experiments, and (3) establishing a measurement framework with corrected metrics (true JSD, task-specific effect sizes) that avoids the pitfalls of flattened distributional comparison.

5.2 Corrected Metrics

Two methodological corrections substantially improve the measurement framework:

True JSD vs. JS distance. The original implementation reported JS distance (the square root of divergence) as "JSD." Since $d_{\mathrm{JS}} = \sqrt{D_{\mathrm{JS}}}$, the numerical values of true divergence are smaller (squared), but the qualitative U-shape is preserved.

Task-specific alignment. For multi-variable tasks (sunk cost, overconfidence, anchoring), computing JSD on flattened arrays conflates the stimulus variable with the response variable. For example, in the sunk cost task, the investment level is uniform by construction in both human and LLM simulations, so a flattened JSD is dominated by this shared component. Our revised metrics evaluate alignment only on the decision-relevant variable or on the task-specific effect size.

5.3 Implications for LLM-Based Behavioral Simulation

If the non-monotonic pattern holds empirically, it implies that the most capable model may not be the best proxy for human decision-making. Practitioners designing LLM-based behavioral simulations should consider calibrating the reasoning mode—e.g., via prompting strategy, decoding temperature, or chain-of-thought depth—to match the target population's behavioral profile rather than maximizing raw capability.

5.4 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. The Gaussian alignment function is one of many possible choices; alternative shapes (e.g., sigmoid, piecewise linear) could produce different quantitative predictions. Statistical power is limited by $n = 10$ capability levels for correlation analyses. Future work should validate these patterns using actual LLM APIs across model families and scales, and extend to additional decision tasks and cultural contexts.

6 CONCLUSION

We have presented a mechanistic simulation framework demonstrating a non-monotonic relationship between reasoning capability and human-like decision fidelity, with alignment peaking at intermediate reasoning depth. The framework uses corrected Jensen-Shannon divergence (true divergence, not distance) and task-specific effect-size metrics to avoid measurement artifacts present in prior flattened-distribution approaches. While the U-shape is built into the model via the Gaussian alignment function, the framework generates concrete, testable predictions and provides a measurement toolkit for future empirical work comparing real LLM outputs to human behavioral baselines.

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] Dominik M Endres and Johannes E Schindelin. 2003. A New Metric for Probability Distributions. *IEEE Transactions on Information Theory* 49, 7 (2003), 1858–1860.

[5] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to Improve Bayesian Reasoning without Instruction: Frequency Formats. *Psychological Review* 102, 4 (1995), 684–704.

[6] Dale Griffin and Amos Tversky. 1992. The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology* 24, 3 (1992), 411–435.

[7] 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.

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

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

[10] 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.

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

[12] 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).

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

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

[15] 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.