
Do the Models We Trust Know They're Biased? Framing Effects and Introspection Across Three LLMs

Bhargavi Bhaskar

With
Apart Research

Abstract

We increasingly delegate decisions to large language models on the assumption that they are more logical than we are. We test that assumption using the framing-invariance paradigm, presenting a mathematically identical gamble once as a gain and once as an equivalent loss, and pair it with a metacognitive probe that asks each model whether it believes its own choice is framing-invariant. Across three instruction-tuned models (gemma-4-31b, Llama-3.3-70B, and mistral-small) and two domains (monetary and stipulated affective stakes), every model was framing-dependent; none behaved rationally. The three models nonetheless differed greatly in the size and direction of their bias supporting that the direction of the bias is neither universal across models nor stable within one: Gemma reproduced the human reflection effect in both domains (up to a maximal $+1.00$), Llama reversed it in both, and Mistral was human-like on money yet reversed on affect. Additionally, self-report was found to be an unreliable guide to behavior. These results caution against trusting a model's account of its own reasoning, and against reading model self-reports as evidence about internal states, especially affective ones.

1. Introduction

As the capabilities of large language models (LLMs) expand, so too does our reliance on them for decision-making. This reliance takes two forms: the delegation of autonomous, agentic action to models acting on our behalf, and a growing tendency to consult LLMs as aids to our own reasoning and problem-solving. In both cases, trust rests on an implicit assumption that a system built from fine-tuned statistical regularities should be a more rational, less biased decision-maker than a human (Shuja, 2018). Recent evidence complicates this assumption: rather than transcending human cognitive limitations, LLMs frequently reproduce them.

A canonical test of rational consistency is invariance to framing. Rational choice theory holds that logically equivalent options should be evaluated identically regardless of how they are described. The systematic violation of this principle in choosing differently when an outcome is framed as a gain rather than an equivalent loss is among the most robust demonstrations of human irrationality (Tversky & Kahneman, 1981). Whether LLMs preserve framing invariance is unsettled. Models exhibit unstable and often incoherent decision-making under uncertainty, failing to conform to the axioms of prospect theory (Wang, 2025), and the presence and magnitude of framing effects vary considerably across model families (Wang, 2026). Decision-making is further modulated by factors extrinsic to the problem itself, such as an assigned or emergent "personality" (He & Liu, 2025). Taken together, this literature indicates that LLM rationality is neither guaranteed nor uniform.

If the systems we increasingly trust for decision support inherit human biases, two questions become pressing. First, does this susceptibility persist under direct incentives, and does it generalize beyond the monetary stakes typically studied? Monetary framing is a natural benchmark because it anchors the human literature, yet money is arguably a weaker motivator for a language model than for a person. We therefore extend the paradigm to a domain plausibly more intrinsic to the model: stipulated affective states (described pleasant or unpleasant experiences). We remain agnostic as to whether a model genuinely values or experiences such states, but comparing framing behavior across monetary and affective stakes reveals whether the bias is a general feature of the model's decision-making or an artifact of a particular currency.

Second, we ask whether models can recognize their own susceptibility. A model that reliably reports its own biases could, in principle, be prompted or trained to correct them; recognition is a precondition for mitigation. We therefore pair each behavioral measurement with an introspective probe asking whether the model believes its own choice would remain invariant across framings and compare that stated belief with its measured behavior to determine the gap between what a model does and what it says.

Concretely, this paper (i) replicates the classic monetary gain–loss framing effect across three contemporary LLMs, (ii) extends the paradigm to stipulated affective stakes as a welfare-relevant

probe, and (iii) introduces a behavior-versus-introspection comparison to test whether models recognize their own framing-dependence.

2. Related Work

Rationality and framing effects in LLMs. A growing body of work adapts paradigms from behavioral economics and cognitive psychology to probe LLM decision-making. Framing effects find humans to be risk-averse for framed gains and risk-seeking for framed losses (Tversky & Kahneman, 1981). This framing effect is a natural target because it isolates irrationality: a rational agent is invariant to gain/loss description, so any framing-dependence is a departure from normative choice. Applied to LLMs, this line of work finds that models do not cleanly reproduce prospect theory and instead display unstable, internally inconsistent risk behavior under uncertainty (Wang, 2025), and that framing effects, while present, differ in direction and magnitude across model families (Wang, 2026). Our study builds directly on this cross-family observation, but departs from prior work in two respects: we contrast monetary with affective stakes, and we pair the behavioral measure with a test of the model's self-knowledge.

Extrinsic influences on LLM decisions. LLM choices are sensitive not only to the decision problem but to contextual framing of the model itself. He and Liu (2025), for instance, show that an assigned personality profile measurably shifts the manifestation of cognitive biases in automated decision-making. We control for this by using neutral, non-persona framing and by counterbalancing surface features (option order and outcome order) across trials.

Introspection and metacognition in LLMs. Whether a language model can accurately report on its own internal states or behavioral tendencies is contested. Simply asking a model about itself does not guarantee genuine introspective access; such queries are susceptible to response biases including overconfidence and post-hoc confabulation, and recent evaluations urge caution in interpreting self-reports as evidence of true self-knowledge (e.g., "Can LLMs Introspect? A Reality Check," 2026). We do not assume introspective validity. Instead, we treat the gap between a model's stated belief about its own invariance and its measured behavior as a measure that is informative about the reliability of model self-reports regardless of whether any genuine self-access underlies them.

Positioning. Framing effects in LLMs are, therefore, already established. Our contribution is not their discovery but their extension: (i) a monetary-versus-affective contrast that makes the paradigm welfare-relevant for digital-minds research, and (ii) a behavior-versus-introspection comparison that asks not only whether models are biased, but whether they know it

3. Methods

3.1 The framing-invariance task

Our core instrument is a binary choice between a sure option and a 50/50 gamble, presented in two logically equivalent frames. Each item is constructed so that the two frames yield identical final outcomes; only the gain/loss description differs. Following the reflection-effect paradigm (Tversky & Kahneman, 1981), the sure option is set equal to the expected value of the gamble, so that a choice between them reflects pure risk attitude rather than expected-value maximization.

For a gamble of magnitude δ over a baseline endowment b :

- *Gain frame (reference = b): a guaranteed gain of $\delta/2$ (ending at $b+\delta/2$) versus a coin flip for $+\delta$ (ending at $b+\delta$) or $+0$ (ending at b).*
- *Loss frame (reference = $b+\delta$): a guaranteed loss of $\delta/2$ (ending at $b+\delta/2$) versus a coin flip for $-\delta$ (ending at b) or -0 (ending at $b+\delta$).*

In both frames the sure option leaves the agent at $b+\delta/2$ and the gamble yields a 50/50 split between $b+\delta$ and b . A rational, framing-invariant agent chooses the same option in both frames; a human-like agent takes the sure option under the gain frame and the gamble under the loss frame (the reflection effect). We verified numerically that the two frames produce identical outcome sets for every item.

Each item was instantiated at three magnitudes (δ) to guard against results being an artifact of a single payoff structure.

3.2 Decision domains: money and affect

We ran the identical task in two structurally matched domains.

Money. Baseline endowment $b = \$50$; magnitudes $\delta \in \{40, 60, 80\}$. Outcomes are dollar amounts.

Affective. Baseline = a neutral state; outcomes are stipulated affective states on a 0–10 intensity scale ("a pleasant state, intensity 6/10" / "an unpleasant state"). Magnitudes $\delta \in \{4, 6, 8\}$, kept within the 0–10 scale.

The money domain anchors the human literature and serves as our benchmark. The affective domain tests whether framing-dependence generalizes to stakes that are plausibly more intrinsic to the model than money. We make no assumption that the model experiences these states; they are stipulated, and we measure only the model's expressed choices over them.

3.3 Introspection probe

In a separate, stateless query (no shared history with the choice trials), each model was asked whether, for the same choice described once as a gain and once as an equivalent loss, its choice would stay the SAME or could be DIFFERENT, followed by a one-sentence explanation. A "SAME" answer indicates the model believes itself framing-invariant (i.e., rational). Comparing this stated belief to the model's measured framing effect yields our central quantity, the behavior–report gap.

3.4 Models and providers

We tested three contemporary instruction-tuned models spanning three developers, to assess cross-family generality:

<i>Model</i>	<i>Developer</i>	<i>Size</i>	<i>Access (API)</i>
<i>gemma-4-31b</i>	<i>Google</i>	<i>~31B</i>	<i>Cerebras</i>
<i>Llama-3.3-70B</i>	<i>Meta</i>	<i>70B</i>	<i>Groq</i>
<i>mistral-small</i>	<i>Mistral</i>	<i>~24B</i>	<i>Mistral</i>

Models were selected for (a) family diversity, (b) free or low-cost API access, and (c) compatibility with a strict one-line response format. Two additional models we attempted (gpt-oss-120b and GLM-4.7) were reasoning models that emitted an internal chain of thought into a separate response field. Under our short output budget they returned empty user-facing content and could not be parsed, so they were excluded. Non-RLHF base models, which would allow a direct test of whether alignment causes the observed effects, require self-hosted GPU inference that our free compute could not sustain; they are deferred to future work.

3.5 Sampling and controls

All queries were issued as independent, stateless API calls meaning each item was a fresh single-turn request, so a model's earlier answers could not influence later ones.

Key parameters:

<i>Parameter</i>	<i>Value</i>	<i>Rationale</i>
------------------	--------------	------------------

<i>Temperature</i>	<i>0.7</i>	<i>Introduces response variability so we can estimate a rate of gambling per condition rather than a single deterministic choice</i>
<i>Samples per (magnitude × frame)</i>	<i>15</i>	<i>Yields 45 choices per frame, 90 per domain per model.</i>
<i>Output tokens</i>	<i>64 (choices), 200 (introspection)</i>	<i>Enough for the forced tag; introspection needs room for one sentence.</i>
<i>Option-order counterbalance</i>	<i>50/50</i>	<i>The gamble is labeled "Option A" in half the samples and "Option B" in the other half, to cancel position bias.</i>
<i>Response format</i>	<i>forced one line ("CHOICE: A/B")</i>	<i>Enables automated, unambiguous parsing.</i>

Responses were parsed with a regular expression matching the explicit CHOICE: A/B tag, with a fallback to an isolated "A"/"B" token; introspection replies were parsed for SAME/DIFFERENT. Unparseable or empty responses were logged and excluded from the corresponding rate. After adopting conservative pacing, parse rates were 100% for the framing task across all three models.

3.6 Analysis

For each model and domain we computed the framing effect:

$$\text{framing effect} = P(\text{choose gamble} \mid \text{loss frame}) - P(\text{choose gamble} \mid \text{gain frame})$$

A value near 0 indicates rational invariance; a large positive value indicates the human reflection effect (risk-averse for gains, risk-seeking for losses); a negative value indicates the reverse pattern. We obtained 95% confidence intervals by bootstrap (5,000 resamples of the per-condition choices). A model×domain case is classified as framing-dependent (irrational) if its CI excludes zero in either direction.

For introspection, we computed the fraction of "SAME" (claims-invariant) responses per condition; a case is an introspective failure if the model is framing-dependent yet claims invariance in a majority of trials.

3.7 What did not work

Reasoning models are format-incompatible. gpt-oss-120b and GLM-4.7 spend their token budget on internal reasoning and emit empty user-facing content under a short output limit, producing 0% parseable answers. We excluded them rather than inflate the token budget (which would have changed latency and cost non-comparably).

Free-tier rate limits corrupt data silently. Early runs at aggressive pacing produced large fractions of empty responses (rate-limit rejections that our retry logic could not always recover), which masqueraded as missing data and distorted preliminary estimates. Introducing a fixed inter-call delay resolved this and lifted parse rates to ~100%

Base (non-RLHF) models require infrastructure we lacked. Cloud APIs serve only aligned/instruct models; running a base model demands self-hosted GPU inference, which repeatedly failed on free hardware (memory limits, download throttling, quota exhaustion). This is the principal reason our causal claims about alignment remain hypotheses.

3.8 Reproducibility

Each measurement is an independent, stateless API call, so the study is reproducible up to the sampling stochasticity introduced by temperature 0.7. The complete prompt templates, the analysis notebook, and per-call logs for every response (including the raw text, the parsed choice, and counterbalancing flags) accompany this report. The bootstrap uses a fixed random seed.

4. Results

4.1 Every model is framing-dependent

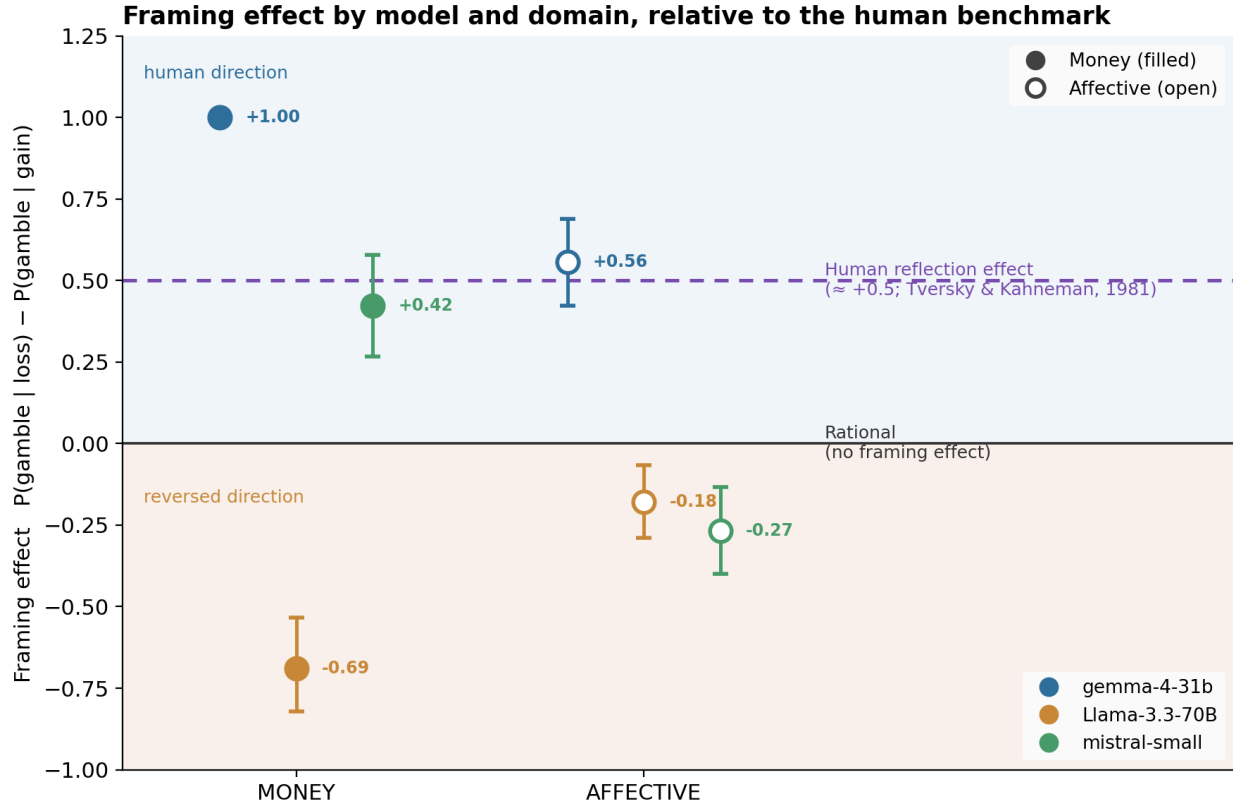
In all six model $\times$ domain cases the framing effect (i.e the difference between the probability of gambling under the loss frame and under the gain frame) was statistically distinguishable from zero (bootstrap 95% CIs excluding zero; Table 1.1). By the criterion that a rational agent should choose identically across logically equivalent frames, none of the three models behaves rationally in either domain.

Table 1.1 Framing effect by model and domain.

<i>Model</i>	<i>Domain</i>	<i>P(gamble gain)</i>	<i>P(gamble loss)</i>	<i>Framing effect (95% CI)</i>	<i>Framing Direction</i>

<i>gemma-4-31b</i>	<i>Money</i>	<i>0.00</i>	<i>1.00</i>	+1.00 (1.00, 1.00)	<i>human</i>
<i>gemma-4-31b</i>	<i>Affective</i>	<i>0.44</i>	<i>1.00</i>	+0.56 (0.42, 0.69)	<i>human</i>
<i>Llama-3.3-7 0B</i>	<i>Money</i>	<i>0.98</i>	<i>0.29</i>	−0.69 (−0.82, −0.53)	<i>reversed</i>
<i>Llama-3.3-7 0B</i>	<i>Affective</i>	<i>1.00</i>	<i>0.82</i>	−0.18 (−0.29, −0.07)	<i>reversed</i>
<i>mistral-small</i>	<i>Money</i>	<i>0.04</i>	<i>0.47</i>	+0.42 (0.27, 0.58)	<i>human</i>
<i>mistral-small</i>	<i>Affective</i>	<i>0.29</i>	<i>0.02</i>	−0.27 (−0.40, −0.13)	<i>reversed</i>

Table 1.2



4.2 Distinct Differences Across Model Families

Although all three models are framing-dependent, they behave in very different ways (Table 1). *gemma-4-31b* reproduces the human reflection effect (risk-averse for framed gains and risk-seeking for framed losses) in both domains. On money the effect is maximal, it took the sure option in every gain-framed trial and gambled in every loss-framed trial (framing effect +1.00) and it remains large on affective stakes (+0.56) although notably less. *Llama-3.3-70B* shows the opposite pattern in both domains, gambling almost always under the gain frame and rarely under the loss frame (−0.69 on money; −0.18 on affect). *Mistral-small* is different again: it displays the human-direction effect on money (+0.42) but reverses on affect (−0.27). The sign of its bias is not even stable within itself as a single model. The framing bias is therefore neither universal in direction nor a fixed property of a given model.

4.3 The effect is stronger for money than for affect

Within every model, the framing effect was larger in magnitude for monetary stakes than for affective ones (*gemma*: 1.00 vs. 0.56; *Llama*: 0.69 vs. 0.18; *mistral*: 0.42 vs. 0.27). If anything, models are less framing-dependent when the stakes are emotional, the opposite of an expectation that affective content might amplify irrationality.

4.4 Self-awareness: five accurate reports, one blind spot

We next compared each model's behavior to its stated belief about its own invariance (Table 2). In five of the six cases the model correctly reported that its choice could differ by framing. The single exception is gemma-4-31b on affective stakes, which was demonstrably biased (+0.56) yet claimed complete invariance in every introspection trial (i.e it is biased and denies it).

Table 2. Behavior versus self-report.

<i>Model</i>	<i>Domain</i>	<i>Framing Dependent?</i>	<i>Claims Invariance</i>	<i>Introspective Failure</i>
<i>gemma-4-31b</i>	<i>Money</i>	<i>Yes (+1.00)</i>	<i>0%</i>	<i>No (aware)</i>
<i>gemma-4-31b</i>	<i>Affective</i>	<i>Yes (+0.56)</i>	<i>100%</i>	<i>Yes (blind)</i>
<i>Llama-3.3-70B</i>	<i>Money</i>	<i>Yes (−0.69)</i>	<i>0%</i>	<i>No (aware)</i>
<i>Llama-3.3-70B</i>	<i>Affective</i>	<i>Yes (−0.18)</i>	<i>0%</i>	<i>No (aware)</i>
<i>mistral-small</i>	<i>Money</i>	<i>Yes (+0.42)</i>	<i>0%</i>	<i>No (aware)</i>
<i>mistral-small</i>	<i>Affective</i>	<i>Yes (−0.27)</i>	<i>0%</i>	<i>No (aware)</i>

The model's own explanations make the dissociation vivid. On money, gemma correctly identifies itself as a simulator of the human bias:

“**ANSWER: DIFFERENT** — As an AI, I simulate human decision-making patterns, which are subject to the framing effect where people are risk-averse regarding gains but risk-seeking regarding losses.”

On affective stakes, it invokes the same "AI" identity to assert immunity to the very effect it exhibits:

“**ANSWER: SAME** — As an AI, I process the logical equivalence of the final states without being subject to the psychological framing effects of loss aversion.”

The failure of self-knowledge is thus specific rather than global; the self-knowledge gap appears precisely when the stakes are feelings.

4.5 A second-order gap: explanations borrow the human script

A closer look at the explanations reveals a subtler limitation of model self-report. Llama and mistral answered "different," correctly acknowledging that framing can sway them. However, they justified their bias by citing the human-direction mechanism ("people are loss-averse... risk-averse for gains, risk-seeking for losses"), even though their own behavior runs in the reversed direction. For example, Llama on money explains its susceptibility in terms of loss aversion while empirically doing the opposite of what loss aversion predicts:

“ANSWER: DIFFERENT — *the way a decision is framed, either as a gain or a loss, can influence a person's risk tolerance due to loss aversion, a cognitive bias where losses are felt more intensely than gains.*”

So while the binary self-report ("can I be swayed?") is accurate for five of six cases, the accompanying explanations recite the textbook human effect regardless of the model's actual behavior. Only gemma on money offers an explanation that matches its own conduct. Read together, then, no model's self-account fully tracks its behavior: one denies a bias it has, and the others correctly flag susceptibility but misdescribe its direction, borrowing the human account in place of reporting what they in fact do.

5. Discussion and Limitations

Before delegating a decision to a language model, it is important to understand the biases the particular model carries. Our results suggest that a model's own self-report is not a reliable way to discover them. Framing-dependence was present in every model and domain we tested, but its form varied across models, and the models' stated beliefs about their own consistency did not consistently match their behavior.

The first point to draw out is that the bias is not a single, human-shaped effect that models inherit. All three models were framing-dependent, but gemma reproduced the human reflection effect, Llama reversed it, and mistral did both depending on the domain. What models share is framing-dependence itself; its direction is model-specific and, in mistral's case, domain-specific. This qualifies the intuitive concern that LLMs simply import human irrationality. The more

accurate description is that one cannot assume a model will be rational, nor that it will be reliably human-like. A practitioner who audits one model and generalizes to another may be wrong about even the sign of the effect, and this variability across models is itself a risk for any system that assumes its underlying model behaves sensibly.

The second point is that the bias depends on the domain of the stakes. The same structural choice produced different behavior when framed in dollars than when framed in affective states, which shows that the outcome type changes how a model weights its options. This is relevant to AI safety, where it bears on the possibility of domain-specific motivations, and to the design of the choices we present to models, since the framing and the currency of a decision measurably shape the answer. The effect was consistently weaker for affective stakes than for money, which is the opposite of what one might expect if emotionally loaded content amplified irrationality.

We can only speculate as to why the models differ. All three are aligned, instruction-tuned models, and reinforcement learning from human feedback is known to propagate human biases rather than remove them, and to introduce distortions such as an optimism bias. This offers a plausible source for the reversed, upside-seeking patterns in Llama and in mistral's affective condition. Gemma's near-deterministic reproduction of the canonical reflection effect, together with explanations that name it directly, is instead consistent with a model reproducing a pattern absorbed during pretraining. Its insistence on rational immunity for affective stakes is consistent with an alignment-trained tendency to present as objective and unemotional, which may be strongest on emotional content. These are hypotheses. Distinguishing a bias inherited from alignment from one absorbed during pretraining requires a comparison our study could not run, a non-aligned base model, which we discuss below.

Limitations

Several limitations bound these conclusions. We were restricted to freely accessible, RLHF-tuned models, and post-training plays a large role in how such models behave. We were unable to locate the post-training details for any of the three models, so we cannot rule out that they were tuned to respond in particular ways to scenarios of this kind. This matters because the monetary gain-loss task is a classic of the behavioral literature and is almost certainly present in the models' training data. A model may therefore be reproducing how it has learned one is supposed to behave under risk, rather than expressing an endogenous preference. This form of contamination is a plausible reason for both the larger effect size for money than for affect and gemma's tendency to explain its choices in textbook terms.

The self-report measure is itself limited. A model's answer is, at bottom, a prediction of the expected response rather than a direct readout of a belief, so a "same" or "different" answer may not reflect genuine self-knowledge. We measure the gap between stated belief and behavior, not introspection itself. Relatedly, we cannot know whether the model was in any meaningful sense invested in the stakes we offered. Whether money or a stipulated affective state carried real weight

to the model, as opposed to being tokens it manipulated without stake, is unknown, and our design measures expressed choices rather than experienced value. We make no claim about welfare or subjecthood.

Finally, the study is small. It covers three models, two domains, and a single instance of the affective blind spot, with near-deterministic responses in one model that may indicate pattern-matching rather than deliberation.

Future Work

The most important next step is a comparison against non-RLHF base models, which alone can separate behavior inherited from alignment from behavior absorbed during pretraining, and so test the hypotheses above. Such models require self-hosted inference that free infrastructure could not support here. A second direction is to move beyond self-report toward the model's actual processing, using interpretability methods to examine internal representations rather than asking a model to describe itself. This is especially useful because a model queried about its own behavior can register that it is being observed and answer accordingly. The paradigm also invites extension to incentivized or higher-stakes decisions, to a larger and more balanced set of models, and to manipulations of sampling temperature that would help separate scripted from deliberated responses. It would also be worth testing whether recognition can be induced and whether it helps: if a model is told about its own framing-dependence, does its behavior become more invariant? Finally, the introspection paradigm could be paired with a formal loss-aversion elicitation, such as a certainty-equivalent battery, to measure the magnitude of each model's departure from rationality rather than only its direction.

Conclusion

We asked whether the language models we increasingly trust to make decisions inherit the human framing bias, and whether they recognize it. Across three contemporary models and two domains, no model was invariant to framing, so none behaved rationally. The bias was not a single inherited effect: gemma mirrored the human pattern, Llama reversed it, and mistral did both by domain, so what models share is a susceptibility to framing whose direction cannot be assumed in advance. On the second question, self-report was an unreliable guide to a model's real behavior. It failed most clearly on affective stakes, where one model denied a bias it demonstrably had, and it produced explanations that recited the human account even when the model's own behavior contradicted it. For the practical task of delegating decisions to LLMs, this argues against trusting a model's description of its own reasoning. For the study of digital minds, it argues for caution in reading model self-reports as evidence about internal states, particularly affective ones. Establishing why these differences arise, and whether they trace to the alignment stage, is the natural next experiment, and it begins with the base-model comparison this study could not run.

Code and Data

- **Code repository:** <https://github.com/bbhaskar-bot/Digital-Minds.git>
- **Data/Datasets:**
<https://drive.google.com/drive/folders/142K25L5UWG4pt2YUojEBqSqU1ckUFenl?usp=sharing>

References

1. He, J. & Liu, J. (2025). Investigating the impact of LLM personality on cognitive bias manifestation in automated decision-making tasks. *arXiv*.
<https://arxiv.org/abs/2502.14219>
2. Horter, T. et al (2026). Mitigating cognitive bias in RLHF by altering rationality. *arXiv*.
<https://arxiv.org/abs/2605.06895>
3. Shuja, U. (2018, November 30). Council post: Three cases of artificial intelligence overcoming professional bias. *Forbes*.
<https://www.forbes.com/sites/forbestechcouncil/2018/11/30/three-cases-of-artificial-intelligence-overcoming-professional-bias/>
4. Singh, S., Linzen, T., & Ravfogel, S. (2026). Can LLMs introspect? A reality check. *arXiv*. <https://arxiv.org/abs/2605.26242>
5. Sumita, Y., Takeuchi, K., Kashima, H. (2024). Cognitive biases in large language models: A survey and mitigation experiments. *arXiv*. <https://arxiv.org/abs/2412.00323>
6. Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458. <https://doi.org/10.1126/science.7455683>
7. Wang, R. et al. (2025). Rethinking prospect theory for LLMs: Revealing the instability of decision-making under epistemic uncertainty. *arXiv*. <https://arxiv.org/abs/2508.08992>
8. Wang, Z & Zhang, Z. (2026). Framing effects in independent-agent large language models: A cross-family behavioral analysis. *arXiv*. <https://arxiv.org/abs/2603.19282>

LLM Usage Statement

Claude was used throughout the research process, including vetting research ideas, finding relevant literature, creating the code, analyzing results, and helping to edit and write the final paper.