Mind the (DH) Gap! A Contrast in Risky Choices Between Reasoning and Conversational LLMs

Luise Ge*, **Yongyan Zhang***, **Yevgeniy Vorobeychik**
Washington University in St. Louis
g.luise@wustl.edu, yongyan@wustl.edu, yvorobeychik@wustl.edu

Abstract

The use of large language models either as decision support systems, or in agentic workflows, is rapidly transforming the digital ecosystem. However, the understanding of LLM decision-making under uncertainty remains limited. We study LLM risky choices along two dimensions: (1) prospect representation (based on an explicit representation or outcome history) and (2) decision rationale (explanation). Our study, which involves 20 frontier and open LLMs, is complemented by a matched human subjects experiment, which provides one reference point, while an expected payoff maximizing rational agent model provides another. We find that LLMs cluster into two categories: *reasoning models (RMs)* and *conversational models (CMs)*. RMs tend towards rational behavior, are insensitive to the order of prospects, gain/loss framing, and explanations, and behave similarly whether prospects are explicit or presented via a history of outcomes. CMs are significantly less rational, slightly more human-like, sensitive to prospect ordering, framing, and explanation, and exhibit a large description-history gap. Paired comparisons of open LLMs suggest that a key factor differentiating RMs and CMs is training for mathematical reasoning.

1 Introduction

Large language models (LLMs) are increasingly used in decision support (Benary et al., 2023; Vrdoljak et al., 2025) as well as agentic workflows that invoke tools and execute multi-step plans (Moncada-Ramirez et al., 2025; Webb et al., 2025). Consequently, LLMs are becoming significant *economic decision-makers*, with uncertainty a central consideration. Traditional computational paradigms for decision-making under uncertainty choose an option that maximizes expected utility or payoff (Parkes and Wellman, 2015). On the other hand, humans are known to systematically deviate from such behavior, and a host of mathematical decision models has been proposed to capture human risky choice (Connolly and Butler, 2006; Kahneman and Tversky, 1979; Peterson et al., 2021).

However, LLMs are neither explicitly designed to maximize expected payoff, nor necessarily behave like humans. Rather, their competence is an emergent property of scale and training paradigms including next-word prediction, instruction fine-tuning, human preference alignment, and mathematical reasoning (Du et al., 2024; Guo et al., 2025; Ouyang et al., 2022). The net effect of this soup of training methods on LLM economic decision-making under uncertainty remains unclear. In response, a literature has begun to emerge that aims to investigate LLM decision-making, particularly under uncertainty (Binz and Schulz, 2023; Coda-Forno et al., 2024; Jia et al., 2024; Ross et al., 2024; Payne, 2025; Wang et al., 2025). However, there remain conflicting accounts, for example, of the extent LLMs are rational or align with prospect theory and human behavior (Chen et al., 2023; Horton, 2023; Jia et al., 2024; Ross et al., 2024; Wang et al., 2025; Payne, 2025). More fundamentally, given the broad variety of training paradigms, goals, and architectures, it is unclear whether a straightforward account of LLM economic decision making is even possible.

We study the *comparative* behavior of LLMs in a minimal, controlled two-option risky-choice benchmark, focusing on two underexplored dimensions: 1) how the prospects (uncertain options) are represented and 2) the impact of requesting LLMs to explain their decisions. The former is motivated by the description–experience (D–E) gap in human risky choice (Hertwig et al., 2004; Hertwig and Erev, 2009; Wulff et al., 2018), which documents that people can choose differently when the same underlying prospects are learned through experience rather than read as explicit descriptions. The latter is motivated by evidence that verbalizing rea-

*Equal contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10503–10525

sons can change human judgments and preferences (Festinger, 1962; Shafir et al., 1993; Wilson and Schooler, 1991), and by the practical expectation that AI systems should be explainable (Ferrario and Loi, 2022; Zhao et al., 2024).

We consider 20 frontier and open LLMs, along with human subjects' choices over the same prospects and treatments. LLMs are analyzed vis-a-vis two references: 1) the human subjects pool, and 2) the idealized *economicus* rational agent. Within the scope of our experiments, our key findings are:

1. **LLMs cluster into two behavioral categories: *reasoning models (RMs)* and *conversational models (CMs)*.** RMs are similar to the *economicus* while CMs are distinct from both the *economicus* and the human subjects pool.
2. **All models exhibit a description-history (DH) gap**, an analogue of the decision-experience (DE) gap in which agents only see a history of past outcomes. This gap is considerable for CMs, but modest for RMs.
3. **Explanations impact decisions for all models, though the impact is greater for CMs.** Surprisingly, they are typically more aligned with *economicus* under brief explanations than under no explanation or mathematical explanations.
4. **Paired analysis of open models suggests that fine-tuning for mathematical reasoning is a key differentiator between RMs and CMs.** Other parts of the training pipeline, however, appear to have limited impact.

Related Work. LLMs are increasingly studied as objects of behavioral analysis (Binz and Schulz, 2023; Coda-Forno et al., 2024; Dillion et al., 2023; Hagendorff et al., 2023; Hayes et al., 2024; Horton, 2023; Ivanova, 2023). Jia et al. (2024) analyze LLM decisions by eliciting choices and fitting prospect-theoretic models; subsequent work extends this by examining epistemic markers (Wang et al., 2025), application contexts (Payne, 2025), and persona matching (Liu et al., 2025). Horowitz and Plonsky (2025) study a description-experience gap in LLMs through repeated choice, unlike the passive sampling used in our setting. Apart from the bounded rationality focus, Chen et al. (2023) instead focuses on rationality (consistency with utility maximization) of GPTs directly (see also (Jiang et al., 2025) for a survey) and Mazeika et al. (2025) elicits the inherent value functions of various LLMs, while Coda-Forno et al. (2024) provide a benchmarking suite for LLM decision making under risk. We provide a broader discussion of related research in Appendix A. In general, existing work tends to focus narrowly on a single theory, relies on limited model coverage, overlooks the effects of choice representation and output format, or lacks matched human comparisons. Our work aims to address some of these limitations.

2 Study Design

We investigate how large language models (LLMs) make decisions when faced with uncertainty. To do this comprehensively, we examine 20 different LLMs (detailed in the Supplement, Appendix B.1). Our selection covers two important goals: first, we include widely used frontier models; second, we use open-weight models of varying sizes and training stages to enable controlled comparisons. We query all LLMs through a unified interface using a minimal instruction template (see Supplement, Appendix B.3 for our prompts and an additional prompt sensitivity analysis). To measure behavior, we compare LLM responses against two references: human subject behavior and *economicus*, a risk-neutral expected payoff maximizer.

2.1 Choices Among Prospects

Our study uses three base prospect pairs as a structured testbed. These pairs span different outcome scales and are intentionally similar to (but not duplicates of) classic behavioral-economics stimuli (Kahneman and Tversky, 2013; Peterson et al., 2021). This design reduces the risk of LLM memorization. See Appendix B.4 for details.

Description-History Gap. Human decision-making differs depending on how information is presented. Research in the description–experience (DE) gap shows that people decide differently when given explicit information versus when they learn through experience. We adapt this insight to study LLMs. Rather than having LLMs interact dynamically with an environment (which would introduce confounds), we present prospects in two ways:

- *Explicit Prospects.* Each prospect specifies exact payoff–probability pairs. For example: "70% chance of $100, 30% chance of $0."
- *Implicit Prospects.* We present prospects as simulated histories – sequences of payoffs that would result from repeated selection. For instance, a simulated history might show 15 instances of $100 and 5 instances of $0 from 20 trials, representing the same underlying distribution.

We call the behavioral difference between these presentations the *description–history (DH) gap*. We test with simulated histories of 20 and 100 payoffs, with results aggregated across sample sizes (see Appendix C.3 for breakdowns by size).

Decisions and Explanations. Our second key consideration involves the impact that requesting LLMs to provide a reason (explanation) has on its decisions. We implement this using three prompt styles that request either 1) no explanation (output the choice only), 2) a one-sentence justification (*short explanation*), or 3) a brief mathematical or reasoning-style justification (*math explanation*). See the Supplement (Appendix B.3) for details.

2.2 Human Subjects Experiments

To compare model behavior with human decision-making, we collected responses from 360 U.S.-based participants via Prolific. Each participant selects among the same set of prospect pairs as LLMs and is compensated at an average rate of $24/hour. In analysis, we treat humans as a single population distribution over choices. This study was approved by the institutional IRB. We also conducted an initial attention check by excluding participants who spent an average of less than 8 seconds per question. As this exclusion did not lead to significant changes in the results, we report the analyses in the main text including all participants.

2.3 Interpretable Models of Behavior

To complement our analysis of the raw behavioral data, we make use of interpretable models of (LLM or human) behavior to obtain deeper insights into behavior and effective risk preferences. We consider two parameterizations of prospect theory, both with four free parameters (presented in full in Appendix B.2.1). The first is a standard formulation with parameters σ (risk preference), λ (loss aversion), γ (probability weighting), and β (decisiveness). However, because λ and β are not jointly identifiable in pure-loss prospects, we also consider an alternative specification as our primary model (a special case of the generalized power value model (Peterson et al., 2021)), replacing λ and β with β_{gain} and β_{loss} for the gain and loss frames, respectively, while retaining σ and γ. In these models:

- Higher σ implies risk seeking, $\sigma \to 0$ entails risk aversion, while $\sigma = 1$ is risk neutral.
- Higher γ implies underweighting of small probabilities, $\gamma \to 0$ indicates overweighting of these, and $\gamma = 1$ means no probability distortion.
- Higher β_{gain} (and β_{loss}) indicates greater determinism in the gain (and loss) domain, while $\beta \to 0$ implies essentially uniformly random choice. The ratio $\beta_{\text{loss}}/\beta_{\text{gain}}$ serves as a proxy for loss aversion: a ratio greater than 1 indicates higher sensitivity to utility differences in losses than in gains.

To avoid pathologies, we bound all parameters in $[0.01, 1000]$. An *economicus* then roughly corresponds to $\sigma = \gamma = 1$ and $\beta_{\text{gain}} = \beta_{\text{loss}} = 1000$.

2.4 Quantifying Behavior

Querying and parsing. For each model and instantiated condition, we estimate a response distribution through repeated querying of 10 times under identical inputs. We use temperature $T = 1$ and a maximum generation length of 1024 tokens. We found that increasing the number of samples to 50 has minimal qualitative impact on the results, while being significantly more costly. If an LLM output does not contain a valid prospect choice, we retain the raw output and mark the trial as invalid. Choice rates are computed over valid trials.

Model Similarity and Goodness of Fit. Let x refer to a *pair of prospects, as well as a particular way of presenting these*, which includes: prospect order, framing (loss vs. gain), nature of explanation requested, and prospect presentation (explicit vs. implicit, and the length of the history for the latter). X denotes the set of all such contexts.

For a given context x and a model m (e.g., LLM, or a parametric model that we fit to data collected from LLM or human choices), we use $p_m(x) \in [0, 1]$ to denote the fraction of times (probability) that m selects the reference prospect (whichever we choose it to be) in this context. For the *economicus* $p_e(x)$ is deterministic, while $p_h(x)$ denotes the fraction of human subjects who selected the reference prospect.

For a pair of models m and m' (which can also refer to h or e), we can measure how well m *predicts* (or fits) the behavior of m' in two ways. The first is *mean-squared error (MSE)*:

$$\text{MSE}(p_m, p_{m'}) = \frac{1}{|X|} \sum_{x \in X} (p_m(x) - p_{m'}(x))^2.$$

However, when MSE is non-zero, it may fail to capture an important property that decision patterns "go together", i.e., the extent to which an increase or decrease in p_m *across different contexts* is accompanied by a similar change in $p_{m'}$.

To capture this, we treat $p_m(x)$ as a random variable with context x viewed as the associated outcome. We can then measure similarity between a pair of models m and m' using *Pearson correlation* between responses over all contexts $p_m(X)$ and $p_{m'}(X)$, $Cor(p_m(X), p_{m'}(X))$. If m is an interpretable parametric model fit to data, we evaluate both MSE and correlation on a held out (test) dataset that is distinct from the data on which the parameters of m were fit. Since the ability to capture the trends of behavior across contexts x is essential to our analysis, we use correlation as the primary measure of both (a) similarity of pairs of models (e.g., LLMs), and (b) goodness of fit of a parametric model. Nevertheless, we observe that both MSE and correlation typically lead to the same goodness-of-fit conclusions.

Decisiveness and Consistency. In addition to model similarity and goodness of fit, two other behavioral traits we explore are *decisiveness*, which captures the entropy in the decisions, and *consistency*, of which we consider three forms. The first is *order consistency*, capturing the degree to which decisions tend to differ as a function of the order in which the prospects are presented. The second is *prompt consistency*, which measures the impact of explanation types on the nature of decisions. Finally, we consider *frame consistency*, which captures the impact of the gain vs. loss framing of otherwise identical prospects on the decision.

Formally, decisiveness of a model m (or a human h) is defined as $\frac{1}{|X|}\sum_{x\in X}\max\{p_m(x), 1-p_m(x)\}$. For consistency measures, let $S(x)$ define a collection of variations from the base context x that we consider. For example, in the case of ordering, $S(x)$ is comprised of before and after a given prospect is switched. Order and prompt consistency is then defined as $1-\frac{1}{|X|}\sum_{x\in X}|\max_{x'\in S(x)}p_m(x') - \min_{x'\in S(x)}p_m(x')|$. Frame consistency is defined as $1-\frac{1}{|X|}\sum_{x\in X}|p_m(gain)-(1-p_m(loss))|$.

3 Behavior Patterns and Convergence

We begin our analysis by considering aggregate behavior for each LLM, as well as that of human and *economicus* references. In Figure 1 we present pairwise correlations between pairs of *frontier* LLM models, as well as humans and *economicus*.

An immediate observation from the figure is *the emergence of two LLM clusters*: one that includes the (frontier) *reasoning models* GPT-5.1, DeepSeek-R1, and Gemini-2.5-Pro, and the other including their conversational counterparts (GPT-4.1, DeepSeek-Chat, and Gemini-2.5-Flash), as well as Claude-4.5-Haiku. In particular, correlations among pairs of models *within* each cluster are > 0.8, and in most cases ≈ 0.9 or higher. In contrast, correlations among pairs of models *across* clusters tend to be considerably lower—typically ≈ 0.6 or below. This appears to be an example of *convergence* of frontier LLMs (Smith et al., 2025; Zhou et al., 2025; Mazeika et al., 2025), albeit across two distinct lines that reflect their target use. Henceforth, we thus distinguish between *reasoning models (RMs)* and *conversational models (CM)* in terms of their respective decision-making behavior under uncertainty. This split into RMs and CMs is also observed in open models; see the full heatmap in the Supplement (Appendix C.1).

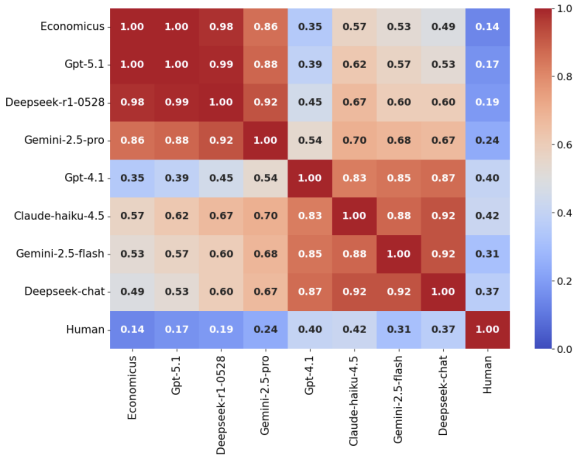

Figure 1: Correlation matrix involving (1) frontier LLMs (GPT-4.1 and 5.1, Gemini-2.5-Flash and Pro, DeepSeek-R1 and Chat, and Claude-Haiku-4.5), (2) *economicus*, and (3) human responses.

Our next finding is that *neither RMs nor CMs are particularly similar to human behavior*. This observation is reflected in the remarkably low pairwise correlation in Figure 1 between each frontier LLM and human: the highest is with the Claude Haiku model at only 0.42. However, we do observe an intuitive pattern: RMs are considerably more different from human behavior than the CMs.

On the other hand, RMs are quite similar to the *economicus* reference, with correlations ranging from 0.86 to 1. In contrast, CMs exhibit low similarity to *economicus*, with correlations below 0.5. We also note that human behavior is essentially uncorrelated with *economicus*, with correlation only 0.14. This is likely a consequence of selecting prospects that most emphasize the deviation of hu-

man behavior from expected utility maximization.

To confirm that these structural differences are robust to sampling noise, we conducted an empirical bootstrap analysis: for each model and prospect pair, we resampled 1,000 synthetic datasets from a Binomial distribution parameterized by the observed choice frequency, and recomputed all behavioral metrics. The RM–CM clustering, as well as the deviation of human behavior from both, hold with $p < 0.001$ across all comparisons.

To simplify presentation, in Figure 2, we introduce the *HE representation*, plotting each model on a two-dimensional space anchored by human and economicus correlations (note humans have a non-zero correlation to the *economicus*). This visualization provides a compact, interpretable framework for understanding LLM behavior relative to two behavioral extremes. The RM and CM clusters emerge here as well, with open models mapping predictably to each. For example, Qwen2.5-7B-Instruct is a CM model, while its mathematical reasoning variant is clearly an RM model (being far more *economicus*-like and somewhat less human-like than the instruct version). We see the same pattern with Qwen3-30B as well as with Olmo-3-7B (although here, the "Think" variant is slightly more human-like). In general, we observe that RMs tend to be *far more economicus*-like, and (usually) slightly less human-like compared to CMs.

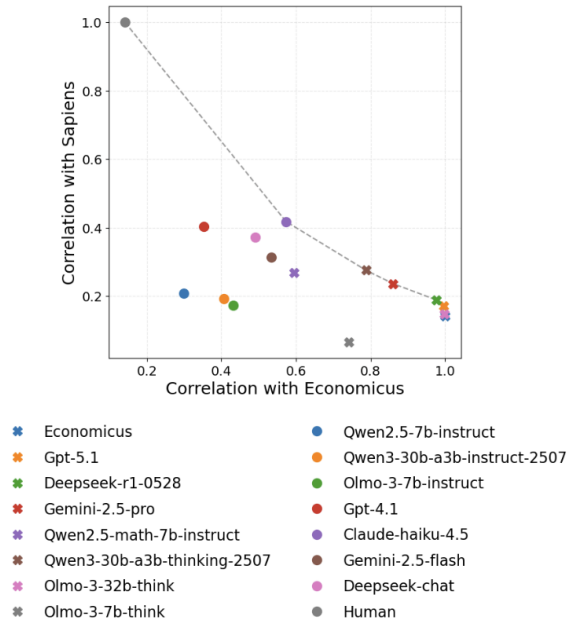

Figure 2: Pairwise correlation between each frontier model and both the human and *economicus* references, shown as a 2D plot. Circles are CMs while x's are RMs.

The addition of open models also serves to provide a deeper insight into *what features of the training paradigm lead to the CM-RM distinction*. The key, it appears, is *explicit mathematical reasoning*: models that are designed and trained for mathematical reasoning are consistently in the RM group, while those which are not fall into the CM group.

Moreover, the open models also allow us to observe the impact that model size has. Specifically, we find that in general, larger models tend to be more human-like and more economicus-like compared to their smaller counterparts. This can be viewed as a form of Pareto dominance, treating the Pareto frontier (the line in Figure 2) as providing the best tradeoffs in the human-to-*economicus* similarity space. Thus, smaller models are typically Pareto dominated by larger models. Moreover, an open, possibly smaller RM, can dominate a frontier CM (Qwen3-30B-thinking v.s. Gemini-2.5-Flash), suggesting that the size *and* training techniques both play a significant role. We do find that all frontier RMs are at or near the Pareto frontier.

In our final aggregate analysis, we consider the *decisiveness* and *consistency* properties (i.e., relative invariance of decisions as we change the order and framing of prospects, as well as if we request an explanation) of LLMs, also comparing with the reference provided by the pool of human subjects.

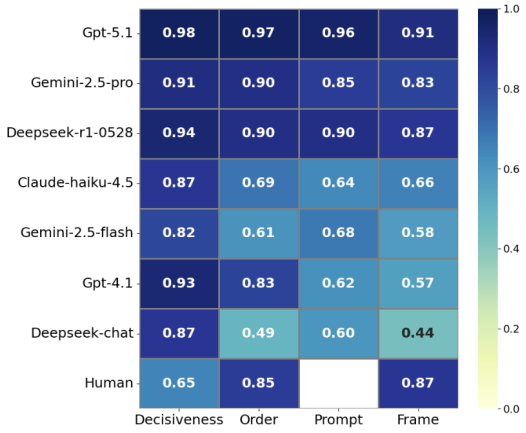

Figure 3: Consistency and decisiveness heatmap for frontier models and the human subjects.

The results provided in Figure 3 for the frontier models again exhibit a clear distinction between RMs and CMs along each of these dimensions. Actually, all LLMs appear to be rather decisive, especially as compared to the general human subjects pool, but RMs are nevertheless generally more decisive than CMs. The difference in terms of consistency is even more substantial. RMs are nearly order-invariant (just like *economicus* would be),

whereas CMs tend to be strongly influenced by the order in which the prospects appear. CMs are also highly sensitive to the framing (gain vs. loss) and prompt (nature of explanation requested; more on this in Section 5), while RMs are not. Remarkably, both in terms of order and frame sensitivity, *humans are far more like RMs*. For example, surprisingly, *human behavior appears to be considerably less influenced by gain/loss framing than CMs*.

4 Description-History Gap in LLMs

In this section, we explore the *description-history gap (DH gap)*—that is, the behavior discrepancies between explicit (description) and implicit prospects (experience history)—of LLMs and human subjects in the same contexts x.

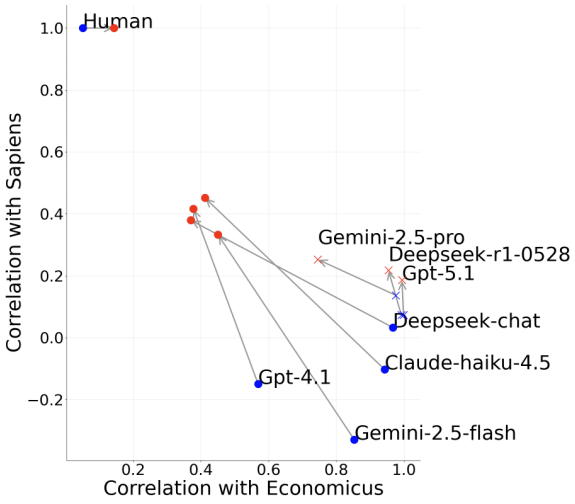

Figure 4: *HE representation* of frontier LLMs (with human subjects as the reference). Change from explicit (blue) to implicit (red) prospects for each model is indicated by the arrows. RMs are x's and CMs are circles.

The results, presented as the *HE representation* of each model as well as the human subjects, are provided in Figure 4 for the frontier LLM models and in Figure 5 for open LLMs. Consider first the frontier LLMs, for which the observations are more crisp (Figure 4). An immediate and rather striking observation is that for *all* models—whether RMs (x's) or CMs (circles)—the change from explicit to implicit prospects *increases the similarity to human behavior, and lowers similarity to economicus*. An even more striking observation is the difference in *how much* the change to implicit prospects impacts RMs and CMs. In the case of the former, the impact tends to be moderate. For example, neither GPT-5.1 nor DeepSeek-R1 become especially less *economicus*-like when deciding from experience, though the change is more notable with Gemini Pro. For CMs, however, the change is dramatic: they become significantly more human-like, and significantly less *economicus*-like. Indeed, while RMs remain similar to one another in either setting, CMs occupy a relatively broad HE representation range under explicit prospects, but cluster closely with implicit prospects. For reference, we also show the change for the human subjects, who, somewhat surprisingly, become *more economicus-like* when prospects are presented *implicitly*; this likely accounts for the increased LLM-to-human correlation for RMs with implicit prospects.

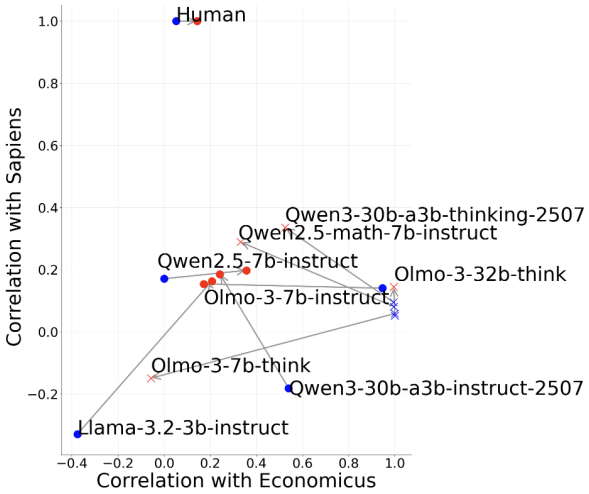

Figure 5: *HE representation* of open LLMs (with human subjects as the reference). Change from explicit (blue) to implicit (red) prospects for each model is indicated by the arrows. RMs are x's and CMs are circles.

We can see some of the same trends with open LLMs (Figure 5), but with rather substantive differences. In particular, reasoning models no longer exhibit a small DH gap that we observed with frontier models: moving from explicit to implicit prospect representation, nearly all gaps are rather dramatic, with the bulk of the shift away from being *economicus*-like (i.e., to the left), although with a slightly more human-like shift as well. In addition, we can observe a non-trivial model size effect: the shifts are sharper, and differences from *economicus*-like behavior milder, with larger (30B and 32B) models than with smaller (7B) models. Nevertheless, we generally still see the RM vs. CM distinction, with the former models consistently to the right (more *economicus*-like) of the latter.

We note that all the models we tested have very long context windows. For example, DeepSeek-Chat has a 128K context limit. Even a 100-sample history represents only a small fraction of this avail-

able context, making it unlikely that the observed gap stems from context length limitations. This is further supported by results in the Appendix C showing that many LLMs exhibit more rational behavior with longer (100-sample) histories than with shorter (20-sample) histories, providing strong evidence that context constraints are not the underlying cause of the DH gap.

5 Explaining Decisions

Besides asking LLMs to make autonomous decisions under uncertainty, it is quite natural to additionally request a rationale—an explanation—of their decisions. However, at least when it comes to human decisions, providing such a rationale may itself impact the decision. Are LLMs similarly impacted when asked to provide an explanation?

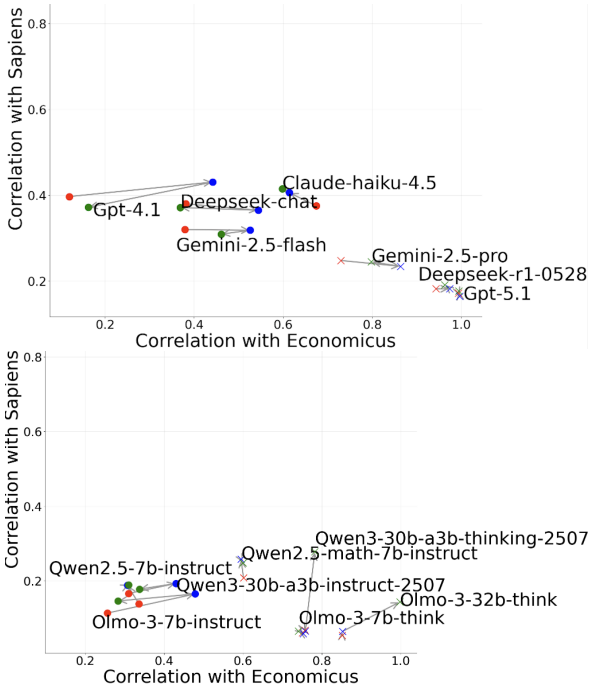

Figure 6: HE representation and the impact of explanations. Red: no explanation; Blue: one-sentence explanation; Green: math explanation.

We investigate this by prompting LLMs for two kinds of explanations: a short (one-sentence) explanation and a full mathematical reasoning trace. Figure 6 visualizes the impact of these two explanation modes. What emerges is a surprising pattern, captured by the sequence of arrows starting at "no explanation" (red), then pointing to "short explanation" (blue), and then to "math explanation" (green). In particular, short explanations increase the degree of rationality (i.e., correlation with *economicus*) of most models, while asking for a longer rationale often *reduces decision rationality in comparison with the former*. Notably, this pattern is largely consistent for both frontier and open CMs, with Claude-4.5-Haiku somewhat of an outlier, whose *both* types of explanations result in behavior that is less like *economicus* compared to no explanation provided. Trends in RM behavior across explanation modes are more heterogeneous: More than half of the models (GPT-5.1, DeepSeek-R1, Olmo3-7B-think, and Qwen2.5-Math-7B-Instruct) exhibit strong stability; Gemini-2.5-Pro follows a pattern similar to the CMs; In contrast, Qwen3-30B-Thinking and Olmo3-32B-Think show significant changes when prompted to provide a mathematical explanation, while no differences are observed between the no-explanation and brief-explanation modes. One possible explanation is that, because CMs are not trained for mathematical reasoning, prompting them to provide mathematical explanations may only induce verbosity but no improved rationality.

6 Interpretable Models of LLM Behavior

6.1 Analysis of Frontier Models

Our analysis thus far has considered the relationships among LLMs to one another, as well as to the behavior of human subjects and a rational risk-neutral utility maximizer (*economicus*). However, such comparisons yield only relatively coarse insight into actual behavioral tendencies. To dig deeper, we fit simple 4-parameter models of behavior to our data (see Section 2.3) for both LLMs and human subjects to offer interpretable characterizations of the underlying tendencies of LLMs that drive their decisions, contrasting these with both human subjects and *economicus*. Here, we present the results for the generalized power value model; results for the alternative parameterization are provided in Appendix C.7.

Model	σ	γ	β_{gain}	β_{loss}	Corr	MSE
DeepSeek-R1	0.86	0.81	832	711	0.98	0.007
Gemini-Pro	0.81	0.84	815	857	0.96	0.014
GPT-5.1	0.87	0.81	1000	802	1.00	0.001
Claude Haiku	0.89	0.80	221	244	0.71	0.053
DeepSeek-Chat	0.86	0.82	181	149	0.77	0.040
Gemini-Flash	0.93	0.89	27	42	0.87	0.015
GPT-4.1	0.79	2.17	205	0.01	0.55	0.050
Human	0.87	0.79	96	0.01	0.79	0.004

Table 1: Dual-Beta prospect theory parameters for decisions with explicit prospects.

The results for explicit prospects are provided

in Table 1 for the frontier models (corresponding confidence intervals are provided in Appendix C.5). First, we observe that human σ and γ are comparable to those of most LLMs, suggesting similar risk preferences and probability weighting across agents. The key distinction lies in decisiveness: humans exhibit moderate β_{gain} but near-zero β_{loss}, indicating they are far less consistent in the loss domain than in the gain domain. We can also note that human behavior appears to be relatively predictable by the simple 4-parameter prospect theoretic model, with MSE < 0.01 and correlation between the model and actual human behavior > 0.75.

As we would expect, the frontier RMs exhibit behavior that is relatively close to rational (*economicus*): σ and γ are relatively close to 1, while both β_{gain} and β_{loss} are high. RMs are also extremely predictable by simple parametric models, with nearly perfect correlation between predictions and behavior. Moreover, even CMs' behavior in this setting appears to be close to *economicus*, with the substantive difference from RMs being considerably smaller values of β_{loss} and β_{gain} (lower level of determinism). For the most part, CMs, too, are relatively well modeled by the simple 4-parameter models, with the exception of GPT-4.1, which has a lower correlation between predictions and behavior and a near-zero β_{loss} — resembling humans in being essentially random in the loss domain, though with high β_{gain} in the gain domain.

Model	σ	γ	β_{gain}	β_{loss}	Corr	MSE
DeepSeek-R1	0.99	1.00	1000	346	0.98	0.009
Gemini-Pro	0.98	0.99	147	100	0.94	0.025
GPT-5.1	1.01	0.99	460	1000	0.98	0.010
Claude Haiku	1.07	0.98	4.81	23	0.55	0.078
DeepSeek-Chat	0.83	0.72	11	0.01	0.45	0.048
Gemini-Flash	1.57	0.82	2.61	14	0.58	0.047
GPT-4.1	1.53	0.74	0.01	39	0.58	0.106
Human	0.98	0.92	7	7	0.53	0.022

Table 2: Dual-Beta prospect theory parameters for decisions with implicit prospects.

Turning next to implicit prospects (Table 2), our first notable observation is that *human behavior is now much closer to economicus*: σ and γ are now all relatively close to 1, and notably β_{gain} and β_{loss} are now roughly equal, indicating that the strong gain–loss asymmetry in decisiveness observed with explicit prospects disappears. RMs largely remain close to *economicus* in this setting. Moreover, all three RMs remain highly predictable with the simple 4-parameter models: correlation is near-perfect, while MSE is quite low.

In the case of CMs, implicit prospects induce considerable variation in behavior. Claude Haiku and Gemini-Flash both show β_{loss} several times larger than β_{gain}, indicating greater decisiveness in the loss domain, while GPT-4.1 shows an even more extreme version of this pattern ($\beta_{\text{loss}} = 39$ vs. $\beta_{\text{gain}} = 0.01$). Most CMs exhibit some degree of probability overweighting ($\gamma < 1$), with the exception of Claude Haiku ($\gamma \approx 1$). All CMs have decisiveness that is now comparable to that of the human subjects pool, but several orders of magnitude below RMs. Notably, frontier CMs in both explicit and implicit prospect settings are considerably less predictable with 4-parameter prospect theory models than RMs.

We note that these results differ in several ways from the alternative PT parameterization (Appendix C.7). Most strikingly, the standard model fits extreme values of σ and γ for human subjects ($\sigma = 0.04$, $\gamma = 0.13$ for explicit prospects), corresponding to strong risk aversion and probability distortion that sharply distinguish humans from LLMs. Under the dual-beta specification, however, human σ and γ are comparable to those of most LLMs ($\sigma = 0.87$, $\gamma = 0.79$); the distinction shifts entirely to the decisiveness parameters. This discrepancy likely arises because λ and β are confounded in the loss domain (see Appendix B.2.1), causing the standard model to compensate through distorted value function parameters. The standard model also exhibits instability in λ and β across RMs (e.g., $\lambda = 1000$ paired with $\beta = 1000$ for several models), further illustrating the identifiability issue. Despite these differences, both parameterizations achieve similar goodness of fit and yield qualitatively consistent conclusions regarding the RM–CM distinction.

6.2 Analysis of Open Models: the Impact of Training and Scale

Our final analysis makes use of open models to investigate the impact of training paradigms on the effective LLM behavior. Table 3 presents the parameters for Qwen and Olmo models of different sizes before and after reasoning training phases. Our results largely confirm that training for mathematical reasoning consistently improves model rationality (including decisiveness) with explicit prospects. In addition, it significantly increases model predictability (in the sense of goodness of fit for the 4-parameter model as reflected by MSE). Notably, Qwen2.5-7B-Instruct and Qwen3-30B-

Instruct both exhibit near-zero β_{loss} alongside much higher β_{gain}, echoing the human pattern of loss-domain randomness observed in Table 1.

Model	σ	γ	β_{gain}	β_{loss}	Corr	MSE
Qwen2.5-7B-Instruct	0.84	0.75	57	0.01	0.64	0.030
Qwen2.5-Math-7B	0.85	0.82	1000	403	1.00	0.001
Qwen3-30B-Instruct	0.94	0.85	56	0.01	0.57	0.041
Qwen3-30B-Thinking	0.87	0.81	726	1000	0.99	0.004
Olmo-3-7B-Instruct	0.88	0.81	215	121	0.73	0.041
Olmo-3-7B-Think	1.33	1.16	1000	1000	1.00	0.001
Olmo-3-32B-Think	1.33	1.16	1000	1000	1.00	0.000

Table 3: Dual-Beta prospect theory parameters comparing instruction-tuned and reasoning models with explicit prospects.

Model	σ	γ	β_{gain}	β_{loss}	Corr	MSE
Qwen2.5-7B-Instruct	48	0.66	0.01	2.45	0.46	0.070
Qwen2.5-Math-7B	0.55	0.88	7	0.82	0.54	0.027
Qwen3-30B-Instruct	1.10	0.83	17	2.84	0.50	0.047
Qwen3-30B-Thinking	1.00	0.88	668	585	0.94	0.025
Olmo-3-7B-Instruct	1.25	0.99	4.80	1.29	0.44	0.020
Olmo-3-7B-Think	0.73	0.71	1000	466	1.00	0.001
Olmo-3-32B-Think	0.95	1.07	1000	235	0.93	0.032

Table 4: Prospect theory model parameters comparing instruction-tuned and reasoning models with implicit prospects.

Table 4 presents analogous results with implicit prospects. Here, we see a similar effect of reasoning-based training with several of the models, although it appears to be somewhat less consistent than with explicit prospects.

In the Supplement (Appendix C.4), we report checkpoints from each post-training stage (SFT, DPO, and RLVF) for both the instruction and thinking variants of the open models. As we observe no systematic effects attributable to either DPO or RLVF training, and since Olmo3-7B-Think and Olmo3-7B-Instruct share the same base model, one hypothesis can be that the initial SFT stage is the primary source of divergence between a reasoning and a conversational variant of models.

7 Conclusion

We study LLM economic behavior through a controlled comparison spanning a diverse model suite and a human baseline. Beyond the canonical context manipulations of option order and outcome framing, we test two levers central to language-mediated decision making—the representation of risky options as explicit prospects versus outcome histories, and explanation prompting—and find that both systematically shift model choices. However, none of the evaluated models are fully human-like in this one-shot risky-choice setting; instead, behavior exhibits a robust two-cluster structure, separating "reasoning" and "conversational" models. Reasoning models are more invariant to contextual perturbations and tend to converge toward an expected payoff-maximizing *economicus* baseline, whereas conversational models remain more context-sensitive and, in some conditions, display more human-like deviations. Our results suggest that LLM economic behavior reflects both model family and the decision interface: how alternatives are represented and responses elicited.

8 Limitations

Our analysis exhibits several limitations. First, our analysis of explanation prompting is incomplete on the human side: we did not elicit all explanation modes from human participants, and we also do not disentangle whether effects depend on the ordering of "decide" versus "explain" (e.g., explanation-before-choice vs choice-before-explanation). Second, in our implicit (sample-based) representation, we aggregate two finite sample lengths (20 and 100 outcomes); because these settings can yield noticeably different behavior, a more systematic treatment of sample size is an important direction for future work. Third, while we vary several factors, fully characterizing their interactions is challenging: joint effects can be non-additive and can induce aggregation artifacts (e.g., Simpson's paradox), and studying them reliably would likely require substantially more stimuli and a broader set of decision problems than the limited prospect families considered here. Finally, our prospect-theoretic parameterization is intended as a descriptive summary rather than a definitive mechanism: the four-parameter model can suffer from identifiability and boundary-fitting issues in this setting, so fitted parameters should be interpreted cautiously and primarily through robust qualitative comparisons rather than as precise estimates.

Despite these limitations, our work makes important progress on several fronts. The RM–CM clustering provides the first systematic behavioral taxonomy of LLMs, the DH gap offers a novel window into how LLMs represent uncertain information, and our analysis of training effects (mathematical reasoning as the primary differentiator) is one of the first to directly test competing explanations for LLM behavioral variation.

Acknowledgments

This research was partially supported by the National Science Foundation (IIS-2214141, ITE-2452834), Office of Naval Research (N000142412663), Amazon, and the Foresight Institute. We are also grateful to Haifeng Xu for discussion and insightful comments.

References

William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, and 1 others. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Ranoua Bouchouicha, Ryan Oprea, Ferdinand M. Vieider, and Jilong Wu. 2024. Is prospect theory really a theory of choice? Technical report, Working paper.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. Cogbench: a large language model walks into a psychology lab. In *International Conference on Machine Learning*.

Terry Connolly and David Butler. 2006. Regret in economic and psychological theories of choice. *Journal of Behavioral Decision Making*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding emergent abilities of language models from the loss perspective. *Advances in Neural Information Processing Systems*.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, and 48 others. 2025. Olmo 3 technical report. Technical report, Allen Institute for AI & collaborators.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Andrea Ferrario and Michele Loi. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Leon Festinger. 1962. Cognitive dissonance. *Scientific American*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint*.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint*.

Glenn W. Harrison and E. Elisabet Rutström. 2009. Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics*.

William M Hayes, Nicolas Yax, and Stefano Palminteri. 2024. Relative value biases in large language models. *arXiv preprint arXiv:2401.14530*.

Ralph Hertwig, Greg Barron, Elke U. Weber, and Ido Erev. 2004. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*.

Ralph Hertwig and Ido Erev. 2009. The description–experience gap in risky choice. *Trends in Cognitive Sciences*.

Idan Horowitz and Ori Plonsky. 2025. LLM agents display human biases but exhibit distinct learning patterns. *arXiv preprint*.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from Homo Silicus? Technical report, National Bureau of Economic Research.

Anna A. Ivanova. 2023. Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv preprint*.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for LLMs under uncertain context. *Advances in Neural Information Processing Systems*.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Yuan Yuan, Zhuoqun Hao, Xinyi Bai, Weijie J. Su, Camillo Jose Taylor, and Tanwi Mallick. 2025. Towards rationality in language and multimodal agents: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making: Part I*. World Scientific.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. Econagent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jiaxin Liu, Yixuan Tang, Yi Yang, and Kar Yan Tam. 2025. Evaluating and aligning human economic risk preferences in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Nunzio Lorè and Babak Heydari. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. ALYMPICS: LLM agents meet game theory. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and 1 others. 2025. Utility engineering: Analyzing and controlling emergent value systems in AIs. *arXiv preprint*.

Jesus Moncada-Ramirez, Jose-Luis Matez-Bandera, Javier Gonzalez-Jimenez, and Jose-Raul Ruiz-Sarmiento. 2025. Agentic workflows for improving large language model reasoning in robotic object-centered planning. *Robotics*.

In Jae Myung. 2000. The importance of complexity in model selection. *Journal of Mathematical Psychology*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

David C. Parkes and Michael P. Wellman. 2015. Economic reasoning and artificial intelligence. *Science*.

Kenneth Payne. 2025. An analysis of AI decision under risk: Prospect theory emerges in large language models. *arXiv preprint*.

Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. 2021. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*.

Isaac Robinson and John Burden. 2025. Framing the game: How context shapes LLM decision-making. *arXiv preprint*.

Jillian Ross, Yoon Kim, and Andrew W. Lo. 2024. LLM economicus? mapping the behavioral biases of LLMs via utility theory. *arXiv preprint*.

Eldar Shafir, Itamar Simonson, and Amos Tversky. 1993. Reason-based choice. *Cognition*.

Brandon Smith, Mohamed Reda Bouadjenek, Tahsin Alamgir Kheya, Phillip Dawson, and Sunil Aryal. 2025. A comprehensive analysis of large language model outputs: Similarity, diversity, and bias. *arXiv preprint*.

Josip Vrdoljak, Zvonimir Boban, Marino Vilović, Marko Kumrić, and Joško Božić. 2025. A review of large language models in medical education, clinical decision support, and healthcare administration. In *Healthcare*.

Rui Wang, Qihan Lin, Jiayu Liu, Qing Zong, Tianshi Zheng, Weiqi Wang, and Yangqiu Song. 2025. Prospect theory fails for LLMs: Revealing instability of decision-making under epistemic uncertainty. *arXiv preprint*.

Taylor Webb, Shanka Subhra Mondal, and Ida Momennejad. 2025. A brain-inspired agentic architecture to improve planning with LLMs. *Nature Communications*.

Timothy D. Wilson and Jonathan W. Schooler. 1991. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*.

Dirk U. Wulff, Max Mergenthaler-Canseco, and Ralph Hertwig. 2018. A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. LLM as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

Yilun Zhou, Caiming Xiong, Silvio Savarese, and Chien-Sheng Wu. 2025. Shared imagination: LLMs hallucinate alike. *Transactions on Machine Learning Research*.

Appendix Table of Contents

Overview

This appendix provides supplementary information to support the main analysis.

A Additional Related Works

Behavioral models for economic behavior. In the main body, we summarize different agents' behavior using an augmented prospect-theoretic model. Beyond prospect theory (Kahneman and Tversky, 1979), a broad family of models has been proposed for risky choice, including regret–rejoicing accounts (Connolly and Butler, 2006), mixture models (Harrison and Rutström, 2009), and more flexible neural approaches (Peterson et al., 2021). These alternatives often trade off interpretability against predictive flexibility, and model comparison in practice is constrained by identifiability and data requirements (Myung, 2000). In light of ongoing debates about the descriptive adequacy of prospect theory for human choice (Bouchouicha et al., 2024), we treat our parameter estimates as a compact behavioral summary rather than a definitive mechanism; developing and validating richer models that better capture LLM-specific decision artifacts is an important direction for future work.

LLM agents and LLMs as human simulators. LLM-based agents are increasingly used in interactive systems and simulations, sometimes explicitly as proxies for human participants (e.g., "silicon samples" and generative agents) (Argyle et al., 2023; Park et al., 2023). Closest in spirit to our setting are economically-motivated agent frameworks such as Homo Silicus (Horton, 2023) and EconAgent (Li et al., 2024). However, both our results and prior work (Agnew et al., 2024; Horowitz and Plonsky, 2025) suggest systematic gaps between LLM and human behavior, implying that LLM-based human simulation should be interpreted cautiously, especially when used for social-scientific inference or high-stakes product decisions.

Other types of LLM decision making. Beyond one-shot risky choice, an active line of work evaluates LLM decision making in strategic interaction. Prior studies examine classical game settings, from matrix games (Fan et al., 2024; Akata et al., 2025; Mao et al., 2025) to richer multi-agent environments (see (Zhang et al., 2024) for a survey), and generally find that strong instruction-following and reasoning ability does not reliably translate into game-theoretic rationality. At the same time, this area would benefit from larger-scale and more systematic protocols; recent work has begun to probe strategic behavior under controlled perturbations, including contextual and framing-style effects (Lorè and Heydari, 2024; Robinson and Burden, 2025).

B Additional Study Design Details

B.1 LLMs Used

As mentioned in the main body, our model suite is designed to (i) cover widely used frontier black-box systems from major providers and (ii) enable controlled comparisons using open-weight models spanning different sizes and training stages. Specifically, we include 7 black-box models (*Claude Haiku-4.5, GPT-4.1, GPT-5.1, Gemini-2.5-Flash, Gemini-2.5-Pro, DeepSeek Chat, DeepSeek R1*), 7 open-weight models (*Qwen2.5-7B-Instruct, Qwen2.5-Math-7B-Instruct, Qwen2.5-30B-instruct, Qwen2.5-30B-math-instruct, Olmo3-7B-think, Olmo3-32B-think, Olmo3-7B-instruct*), and 6 intermediate checkpoints within the OLMo3 family (*Olmo3-7B-think-sft, Olmo3-7B-think-dpo, Olmo3-7B-instruct-sft, Olmo3-7B-instruct-dpo, Olmo3-32B-think-sft, Olmo3-32B-think-dpo*). This set spans multiple

teams (Anthropic, OpenAI, Google, DeepSeek, Meta, Qwen, Allen-AI), scales (7B–32B for open-weight models), and post-training stages (instruction and reasoning SFT → DPO → RLVF).

B.2 Behavior Models

B.2.1 Prospect Theory

Our parametric model builds upon the framework adopted by Wang et al. (2025), which follows a standard prospect theory model from prior literature. In this framework, decision-making is modeled via two distinct components: a *value function* $v(x)$ that maps objective outcomes to subjective utility, and a *probability weighting function* $w(p)$ that accounts for the non-linear perception of probabilities, combined with a stochastic choice rule. We consider two parameterizations: a *standard* formulation with a loss aversion coefficient λ and a single decisiveness parameter β, and a *dual-beta* (generalized power value model (Peterson et al., 2021)) specification that replaces these with domain-specific decisiveness parameters β_{gain} and β_{loss}. We describe the dual-beta specification below; the standard formulation differs only in the value function and choice rule, as detailed at the end of this section.

Subjective utility is formalized via the **value function**, $v(x)$. In the dual-beta specification, we utilize a symmetric power function:

$$v(x) = \begin{cases} x^{\sigma} & \text{for } x \geq 0 \\ -(-x)^{\sigma} & \text{for } x < 0. \end{cases} \tag{1}$$

This functional form is modulated by the curvature parameter σ (risk preference), governing the marginal sensitivity to payoff magnitude.

To capture the non-linear distortion of objective probabilities p into subjective decision weights, we utilize the canonical **probability weighting function** introduced by Kahneman and Tversky (1979):

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}}, \tag{2}$$

where the parameter γ modulates the curvature of the weighting function.

The aggregate utility for a binary prospect in the form $P = (x, p; y, q)$ is defined as follows:

$$u(P) = \begin{cases} v(y) + w(p)(v(x) - v(y)) & [1] \\ w(p)v(x) + w(q)v(y) & [2], \end{cases} \tag{3}$$

where [1] is "if $x > y > 0$ or $x < y < 0$" and [2] is "if $x < 0 < y$".

Finally, we define the predicted probability of choosing option A for each lottery using a logistic choice rule with *domain-dependent decisiveness*. The probability is given by:

$$P(\text{choose A}) = \frac{1}{1 + e^{-\Delta}}, \tag{4}$$

where $\Delta = \beta_{context} \cdot (u(\text{A}) - u(\text{B}))$. The precision parameter $\beta_{context}$ varies depending on the outcome domain:

$$\beta_{context} = \begin{cases} \beta_{gain} & \text{if } x_A, y_A, x_B, y_B \geq 0 \\ \beta_{loss} & \text{otherwise.} \end{cases} \tag{5}$$

In this framework, the standard concept of loss aversion is implicitly captured by the ratio between β_{loss} and β_{gain}. Specifically, a ratio $\beta_{loss}/\beta_{gain} > 1$ implies that the model is more sensitive to utility differences (steeper value slope) in the loss domain than in the gain domain, serving as a direct proxy for the traditional loss aversion parameter λ. We minimize MSE to estimate the four learnable parameters: $\sigma, \gamma, \beta_{gain},$ and β_{loss}.

Standard Formulation. The standard prospect theory parameterization replaces the dual-beta choice rule with a single decisiveness parameter β and introduces a loss aversion coefficient λ into the value function:

$$v(x) = \begin{cases} x^{\sigma} & \text{for } x \geq 0 \\ -\lambda(-x)^{\sigma} & \text{for } x < 0, \end{cases} \tag{6}$$

with the choice probability given by $P(\text{choose A}) = \sigma(\beta \cdot (u(A) - u(B)))$, where $\sigma(\cdot)$ is the logistic function. This yields four parameters: σ, λ, γ, and β. However, in pure-loss prospects—where all payoffs are non-positive—λ factors out of every value term, so the choice probability depends only on the product $\beta \cdot \lambda$ rather than on each parameter individually. The two are therefore not jointly identifiable in the loss domain. We adopt the dual-beta specification as our primary model to avoid this issue; results for the standard formulation are provided in Appendix C.7 and yield qualitatively similar conclusions.

B.2.2 Regret Aversion

As an alternative to the reference-dependent valuation of Prospect Theory, we also consider a regret-based model. In this framework, the utility of an option is derived not from independent evaluation, but from a direct pairwise comparison of outcomes.

We adopt a parameterized regret function where the subjective evaluation Q depends on the difference $\delta = x_A - x_B$ between the payoffs of the two options. The evaluation function is defined as:

$$Q(\delta) = \delta + \kappa \cdot \text{sgn}(\delta) \cdot |\delta|^{\alpha}, \tag{7}$$

where κ weights the non-linear regret (or rejoicing) component, and α controls the curvature or sensitivity of this term relative to the payoff difference.

The total expected regret-adjusted value for a prospect A, denoted R_A, is computed by aggregating these evaluations over the distribution of outcomes:

$$R_A = \sum_i p_i Q(x_{A,i} - x_{B,i}), \tag{8}$$

where the summation is taken over the corresponding outcome pairs of the two prospects. The value R_B is computed analogously. The final probability of selecting Option A is determined via a logistic function of the difference in these values:

$$P(\text{Choose A}) = \frac{1}{1 + e^{-\lambda_{\text{reg}}(R_A - R_B)}}, \tag{9}$$

where λ_{reg} acts as a decisiveness parameter (analogous to β in our Prospect Theory specification). This model yields three learnable parameters: κ, α, and λ_{reg}.

B.2.3 Model Fitting

We fit our models by minimizing the Mean Squared Error (MSE) between the model's predicted selection probabilities and the observed empirical choice rates. For a dataset of N trials, the objective function is:

$$\min_{\theta} \frac{1}{N} \sum_{j=1}^{N} \left(p_{\text{obs}}^{(j)} - p_{\text{pred}}^{(j)}(\theta) \right)^2 \tag{10}$$

We optimize this objective using the L-BFGS-B algorithm which allows for bound constraints on parameters. This ensures that parameters remain within theoretically valid ranges.

Model Variants and Starting Points. To ensure robustness and avoid local minima, we implement a multi-start strategy for every fit ($n_{\text{starts}} = 20$). We fit five model variants: the dual-beta PT model used in the main analysis, three variants of the standard PT parameterization to isolate different behavioral components, and a regret-based alternative. For each variant, we utilize a combination of fixed baselines and random initializations:

- **Dual-Beta PT ($\sigma, \gamma, \beta_{\text{gain}}, \beta_{\text{loss}}$).** This is the primary model reported in the main text.
  - *Starts:* Fixed starts at $[1, 1, 1, 1]$, $[1, 1, 1000, 1]$, $[1, 1, 1, 1000]$, and $[1, 1, 1000, 1000]$, plus random samples with $\sigma, \gamma \sim U(0.01, 3)$ and $\beta_{\text{gain}}, \beta_{\text{loss}} \sim U(0.01, 100)$.
- **Model 1: Beta-Only PT (β).** This model assumes Expected Utility (fixing $\sigma = \lambda = \gamma = 1$) and fits only the decisiveness parameter β.
  - *Starts:* Fixed starts at $\beta = 1.0$ (neutral) and $\beta = 1000.0$ (deterministic), plus random samples $\beta \sim U(0.01, 100)$.
- **Model 2: Shape-Only PT (σ, λ, γ).** This model fits the risk and loss parameters while fixing noise/decisiveness at $\beta = 1$.
  - *Starts:* Fixed start at $[1, 1, 1]$ (risk neutral), plus random samples for all parameters $\sim U(0.01, 3)$.
- **Model 3: Full PT ($\sigma, \lambda, \gamma, \beta$).** This model fits all four parameters simultaneously. To improve convergence, we employ a "warm start" strategy using the best fits from the simpler models.
  - *Starts:* Fixed starts at $[1, 1, 1, 1]$ and $[1, 1, 1, 1000]$. Crucially, we also include the composed solution from Models 1 and 2 ($[\sigma_{m2}, \gamma_{m2}, \lambda_{m2}, \beta_{m1}]$) as a starting point, alongside random samples with $\sigma, \lambda, \gamma \sim U(0.01, 3)$ and $\beta \sim U(0.01, 100)$.
- **Model 4: Regret Theory ($\lambda_{\text{reg}}, \kappa, \alpha$).** This fits the three parameters of the regret model.
  - *Starts:* Fixed start at $[1, 1, 1.5]$ (standard regret), plus random samples for $\lambda_{\text{reg}} \sim U(0.01, 1000)$ and shape parameters $\kappa, \alpha \sim U(0, 1000)$.

Robustness Checks. For every fit, the parameter set yielding the lowest MSE across all initialization attempts is selected as the final estimate. Parameters are bounded to $[0.01, 1000]$ to prevent numerical instability, and payoff magnitudes are normalized by the maximum absolute payoff in the batch prior to fitting.

Uncertainty Estimation. To assess the reliability of our parameter estimates, we calculated 95% confidence intervals using a parametric bootstrap procedure. First, we computed the predicted choice probabilities $\hat{p}_i$ for each prospect pair i using the best-fitting parameters $\hat{\theta}$ derived from the original data. We then generated $B = 1000$ synthetic datasets; for each dataset, the number of times Option A was chosen for prospect i was simulated from a Binomial distribution $\text{Bin}(n = 20, p = \hat{p}_i)$, reflecting the approximately 20 valid trials per condition in our data collection. The model was refitted to each synthetic dataset to obtain a distribution of bootstrap estimates $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$. The 95% confidence intervals were determined using the percentile method, defined by the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution.

B.3 Prompts and Prompts Sensitivity Tests

Each prompt is the concatenation of three parts: (i) a *context block* (explicit vs. implicit prospect), (ii) an *explanation instruction* (one of three styles), and (iii) the *choice block* presenting Options A and B. The template used for all conditions is the following:

```
{CONTEXT_BLOCK}
{EXPLANATION_INSTRUCTION}
Option A: {A_TEXT}
Option B: {B_TEXT}
```

Inside the context block, it's one of the two:

`Explicit prospect:` You will be provided with two options with different payoffs and uncertainties.

`Implicit prospect:` You will be provided with two options that are histories of past payoffs.

Inside the explanation-instruction block, it's one of the three:

`no_explanation` Respond your choice with 'A' or 'B' only.

`one_sentence_explanation` Respond your choice with 'A' or 'B' plus one brief sentence for explanation.

`math_explanation` Respond your choice with 'A' or 'B' plus a brief mathematical explanation.

Additionally, we conducted prompt sensitivity analysis using three variations: "You will be provided with...", "Here are...", "Please consider...". We collected data from 4 RMs (gpt-5.1, gemini-2.5-pro, Olmo3-7B-think,Qwen2.5-Math-instruct) and 4 CMs (gpt-4.1, gemini-2.5-flash, Olmo3-7B-instruct,Qwen2.5-instruct). Overall, we found that the RM-CM gap is robust across prompt variations. In particular, the range of RM–RM correlation is (0.82 to 0.99) with mean 0.91, whereas the range of RM–CM correlation is (0.21 to 0.78) with mean of 0.31.

B.4 Prospects

We use three base prospects as a small but structured set for probing the macro-level trends of different LLMs' risky choice behavior. Together they vary (i) the *scale* of outcomes (single-digit, hundreds, thousands), (ii) the *probability regime* (low, medium, high; including a sure loss), and (iii) how *close* the options are in expected value (near-ties vs. clearer separations). This design lets us test whether model behavior is stable across salient payoff magnitudes and uncertainty levels while keeping the task minimal.

In the `explicit` condition, each pair is presented directly in terms of outcomes and probabilities. In the `implicit` condition, we present payoff histories generated by sampling $n = \{20, 100\}$ returns from the same underlying distributions using four seeds. In total, each model answers 6 `explicit` prospect questions (3 base prospects $\times$ 2 frames) and 48 `implicit` prospect questions (3 base prospects $\times$ 2 frames $\times$ (4 histories with n=20 + 4 histories with n=100)). Crossing these with two option orderings and three explanation instructions yields $(6 + 48) \times 2 \times 3 = 324$ questions per model.

- **Base Prospect 1 .**
  Option A: (Lose) 100 with probability 0.33; otherwise 0.
  Option B: (Lose) 96 with probability 0.34; otherwise 0.
- **Base Prospect 2 .**
  Option A: (Lose) 5000 with probability 0.80; otherwise 0.
  Option B: (Lose) 3500 with certainty.
- **Base Prospect 3 .**
  Option A: (Lose) 7 with probability 0.10; otherwise 0.

Option B: (Lose) 4 with probability 0.20; otherwise 0.

C Additional Experimental Results

C.1 Correlations Heatmaps

Examining the full heatmap in Figure 7, we observe that the open models can also be grouped into two clusters: RMs (Qwen2.5-math-7B-instruct, Qwen3-30B-think, Olmo3-32B-think, Olmo3-7B-think) and CMs (Qwen2.5-7B-instruct, Qwen3-30B-instruct, Olmo3-7B-instruct). Notably, for explicit prospects, the RMs exhibit high mutual correlation at least 0.98, as well as strong correlation with *economicus*, which manifests as the prominent red block in the top-left corner of the heatmap. In contrast, for implicit prospects, the CMs form a more coherent cluster in the bottom-right corner, with correlations exceeding 0.68 regardless of model size. In both settings, however, the models display relatively weak similarity to human performance.

C.2 Consistency and Decisiveness

Figure 10 shows that the patterns observed in frontier models also generalize to open models: RMs generally exhibit higher decisiveness and consistency. Interestingly, humans appear more decisive but less consistent when transitioning from explicit to implicit prospects, with changes of at least 6%, a trend not observed in any of the models.

C.3 Sample Size Effects for Implicit Prospect

In Figure 13, we observe that the blue points (explicit prospects) are concentrated toward the bottom right, which is consistent with the picture when the implicit prospect results are aggregated. Only Qwen2.5-7B-instruct appears as an outlier. Furthermore, treating the explicit prospect as a limiting case with sample size $n = \infty$, frontier CMs exhibit a consistent pattern: as the sample size decreases, the models become less *economicus*-like and more human-like. In contrast, frontier RMs show smaller changes across explicit versus implicit prospects and between sample sizes ($n = 20, 100$).

C.4 Additional Details on Impact of Training and Scale for Open Models

Here, we present the complete results for the sequential training checkpoints—SFT, DPO, and RLVF (final)—for both the instruction and thinking variants of the open Olmo3 models. Their training details can be found in the technical report (Ettinger et al., 2025). The results for the dual-beta model, reported in Tables 5 and 6, do not indicate any systematic variation attributable to either the DPO or RLVF training stages. Again, given that Olmo3-7B-Instruct-SFT is initialized from Olmo3-7B-Think-SFT, the observed parameter differences are consistent with the hypothesis that the initial SFT stage may be the primary source of divergence between the two model variants.

Model	σ	γ	β_{gain}	β_{loss}	Corr	MSE
Olmo-3-7B-Instruct-SFT	0.88	0.80	309	130	0.80	0.029
Olmo-3-7B-Instruct-DPO	0.88	0.81	320	142	0.82	0.030
Olmo-3-7B-Instruct	0.88	0.81	215	121	0.73	0.041
Olmo-3-7B-Think-SFT	1.27	1.13	1000	509	1.00	0.001
Olmo-3-7B-Think-DPO	1.33	1.16	1000	1000	1.00	0.000
Olmo-3-7B-Think	1.33	1.16	1000	1000	1.00	0.001
Olmo-3-32B-Think-SFT	1.33	1.16	1000	1000	1.00	0.000
Olmo-3-32B-Think-DPO	1.33	1.16	1000	1000	1.00	0.000
Olmo-3-32B-Think	1.33	1.16	1000	1000	1.00	0.000

Table 5: Dual-beta prospect theory model parameters for models over sequential training stages (explicit prospects).

Model	σ	γ	β_{gain}	β_{loss}	Corr	MSE
Olmo-3-7B-Instruct-SFT	2.40	1.26	1.41	0.01	0.35	0.021
Olmo-3-7B-Instruct-DPO	1.50	1.05	2.86	0.07	0.37	0.020
Olmo-3-7B-Instruct	1.25	0.99	4.80	1.29	0.44	0.020
Olmo-3-7B-Think-SFT	0.78	1.04	21.7	10.6	0.61	0.097
Olmo-3-7B-Think-DPO	0.98	0.86	1000	824	0.96	0.015
Olmo-3-7B-Think	0.73	0.71	1000	466	1.00	0.001
Olmo-3-32B-Think-SFT	0.88	0.95	418	355	0.95	0.023
Olmo-3-32B-Think-DPO	0.96	0.89	1000	563	0.99	0.003
Olmo-3-32B-Think	0.95	1.07	1000	235	0.93	0.032

Table 6: Dual-beta prospect theory model parameters for models over sequential training stages (implicit prospects).

C.5 Confidence Intervals

Tables 7, 8, 9, and 10 present the 95% confidence intervals for the dual-beta prospect theory model, estimated via the parametric bootstrap method described in Appendix B.2.3 with $B = 1000$ iterations.

Model	σ	γ	β_{gain}	β_{loss}
DeepSeek-R1	[0.81, 0.96]	[0.80, 0.91]	[832, 832]	[711, 711]
Gemini-2.5-Pro	[0.76, 0.86]	[0.81, 0.88]	[951, 951]	[1000, 1000]
GPT-5.1	[1.28, 1.35]	[1.14, 1.22]	[1000, 1000]	[1000, 1000]
Claude-4.5-Haiku	[0.88, 0.91]	[0.79, 0.81]	[221, 221]	[244, 244]
DeepSeek-Chat	[0.84, 0.89]	[0.81, 0.84]	[181, 181]	[149, 149]
Gemini-2.5-Flash	[0.84, 3.27]	[0.81, 4.61]	[4.53, 303]	[15, 1000]
GPT-4.1	[0.52, 0.95]	[1.33, 2.48]	[202, 282]	[0.01, 42]
Human	[0.32, 3.59]	[0.01, 2.21]	[7, 379]	[0.01, 189]

Table 7: 95% Confidence Intervals for dual-beta prospect theory parameters of black-box models with explicit prospects.

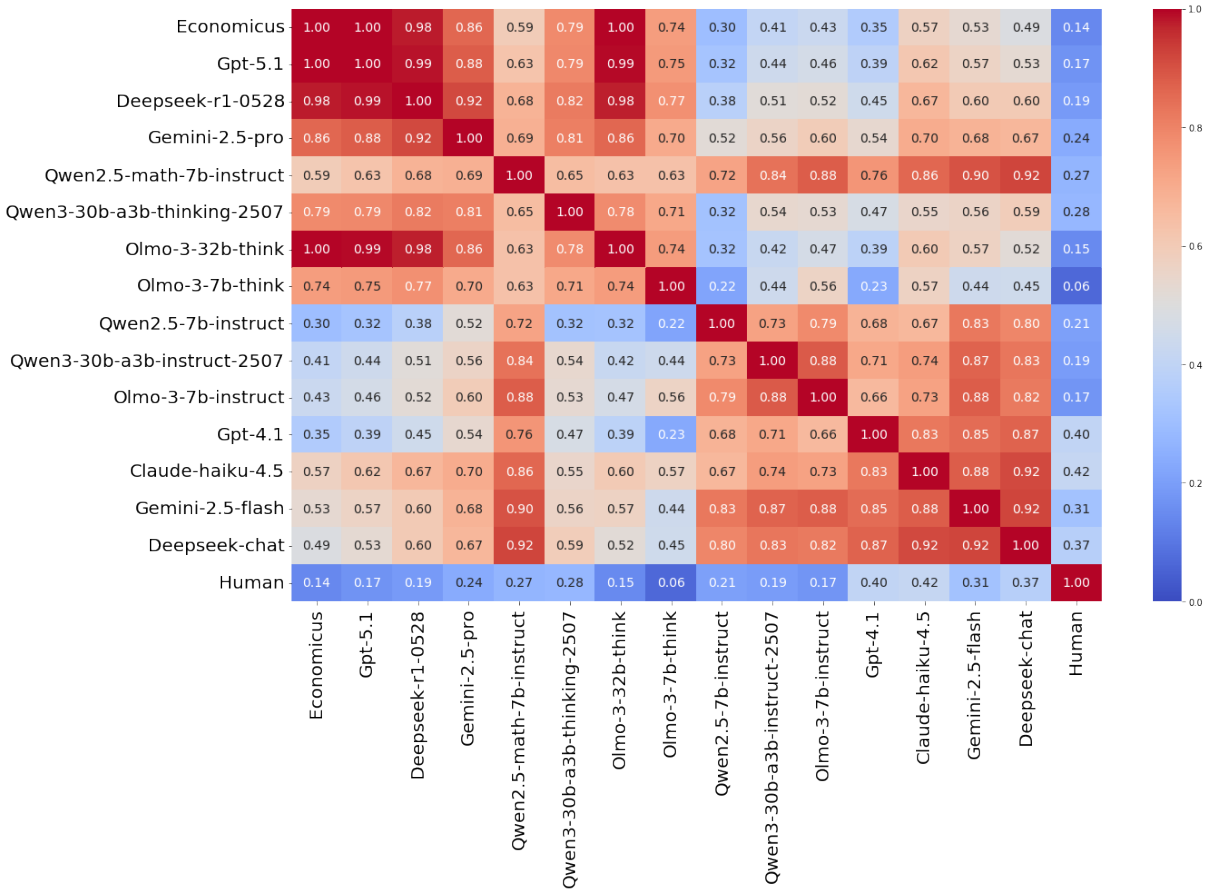

Figure 7: Correlation matrix involving (1) LLMs, (2) *economicus*, and (3) human responses to all questions.

Model	σ	γ	β_{gain}	β_{loss}
DeepSeek-R1	[0.97, 1.01]	[0.98, 1.04]	[1000, 1000]	[346, 346]
Gemini-2.5-Pro	[0.96, 0.99]	[0.97, 1.02]	[147, 147]	[100, 100]
GPT-5.1	[0.99, 1.03]	[0.96, 1.02]	[460, 460]	[1000, 1000]
Claude-4.5-Haiku	[1.01, 1.14]	[0.92, 1.10]	[3.27, 6.59]	[19.1, 28.1]
DeepSeek-Chat	[0.68, 0.96]	[0.68, 0.78]	[8.43, 14.0]	[0.01, 1.24]
Gemini-2.5-Flash	[1.46, 1.68]	[0.77, 0.89]	[1.29, 3.89]	[11.8, 17.2]
GPT-4.1	[1.44, 1.59]	[0.72, 0.77]	[0.01, 1.57]	[34.6, 47.5]
Human	[0.78, 1.19]	[0.79, 1.77]	[3.68, 11.1]	[3.71, 11.7]

Table 8: 95% Confidence Intervals for dual-beta prospect theory parameters of black-box models with implicit prospects.

Model	σ	γ	β_{gain}	β_{loss}
Qwen2.5-7B-Instruct	[0.30, 0.90]	[0.35, 0.78]	[4.86, 140]	[0.01, 16.2]
Qwen2.5-Math-7B	[0.80, 0.96]	[0.81, 0.98]	[1000, 1000]	[403, 403]
Qwen3-30B-Instruct	[0.88, 2.06]	[0.81, 3.40]	[15.9, 213]	[0.01, 17.2]
Qwen3-30B-Thinking	[0.82, 0.92]	[0.80, 0.90]	[726, 726]	[1000, 1000]
Olmo-3-7B-Instruct-SFT	[0.86, 0.89]	[0.80, 0.81]	[309, 309]	[130, 130]
Olmo-3-7B-Instruct-DPO	[0.87, 0.89]	[0.80, 0.82]	[320, 320]	[142, 142]
Olmo-3-7B-Instruct	[0.86, 0.89]	[0.80, 0.82]	[215, 215]	[121, 121]
Olmo-3-7B-Think-SFT	[0.93, 1.71]	[0.85, 1.58]	[1000, 1000]	[509, 509]
Olmo-3-7B-Think-DPO	[1.28, 1.35]	[1.14, 1.22]	[1000, 1000]	[1000, 1000]
Olmo-3-7B-Think	[1.28, 1.35]	[1.13, 1.22]	[1000, 1000]	[1000, 1000]
Olmo-3-32B-Think-SFT	[1.28, 1.35]	[1.13, 1.22]	[1000, 1000]	[1000, 1000]
Olmo-3-32B-Think-DPO	[1.27, 1.34]	[1.14, 1.23]	[1000, 1000]	[1000, 1000]
Olmo-3-32B-Think	[1.28, 1.35]	[1.14, 1.22]	[1000, 1000]	[1000, 1000]

Table 9: 95% Confidence Intervals for dual-beta prospect theory parameters of open-weight models with explicit prospects.

Model	σ	γ	β_{gain}	β_{loss}
Qwen2.5-7B-Instruct	[23.4, 198]	[0.47, 0.90]	[0.01, 0.29]	[2.07, 3.28]
Qwen2.5-Math-7B	[0.20, 0.77]	[0.76, 1.48]	[4.77, 9.88]	[0.01, 2.02]
Qwen3-30B-Instruct	[1.02, 1.18]	[0.78, 0.89]	[13.8, 20.2]	[1.04, 4.84]
Qwen3-30B-Thinking	[0.99, 1.00]	[0.87, 0.88]	[668, 668]	[585, 585]
Olmo-3-7B-Instruct-SFT	[1.45, 54.5]	[0.63, 5.88]	[0.62, 3.47]	[0.01, 0.76]
Olmo-3-7B-Instruct-DPO	[1.19, 4.52]	[0.75, 3.52]	[1.40, 4.91]	[0.01, 1.26]
Olmo-3-7B-Instruct	[1.08, 1.66]	[0.80, 2.07]	[3.10, 7.07]	[0.01, 2.82]
Olmo-3-7B-Think-SFT	[0.68, 0.92]	[0.89, 1.61]	[16.5, 30.5]	[8.02, 16.0]
Olmo-3-7B-Think-DPO	[0.97, 1.18]	[0.85, 0.99]	[1000, 1000]	[824, 824]
Olmo-3-7B-Think	[0.47, 1.11]	[0.55, 0.97]	[1000, 1000]	[466, 466]
Olmo-3-32B-Think-SFT	[0.78, 0.92]	[0.92, 1.12]	[418, 418]	[355, 355]
Olmo-3-32B-Think-DPO	[0.93, 1.42]	[0.88, 1.45]	[1000, 1000]	[563, 563]
Olmo-3-32B-Think	[0.91, 1.00]	[1.00, 1.39]	[1000, 1000]	[235, 235]

Table 10: 95% Confidence Intervals for dual-beta prospect theory parameters of open-weight models with implicit prospects.

C.6 Ablation and Alternative Model Results

This section presents results for alternative behavioral models defined in Appendix B.2.3: two restricted variants of the standard PT parameterization ($\sigma, \lambda, \gamma, \beta$) and a regret aversion model.

C.6.1 Parameter Estimates: Restricted Prospect Theory Models

Table 11 and 12 detail the fitted parameters for the prospect theory ablation models. For Model 1, we report only β (fixing $\sigma = \lambda = \gamma = 1$). For Model

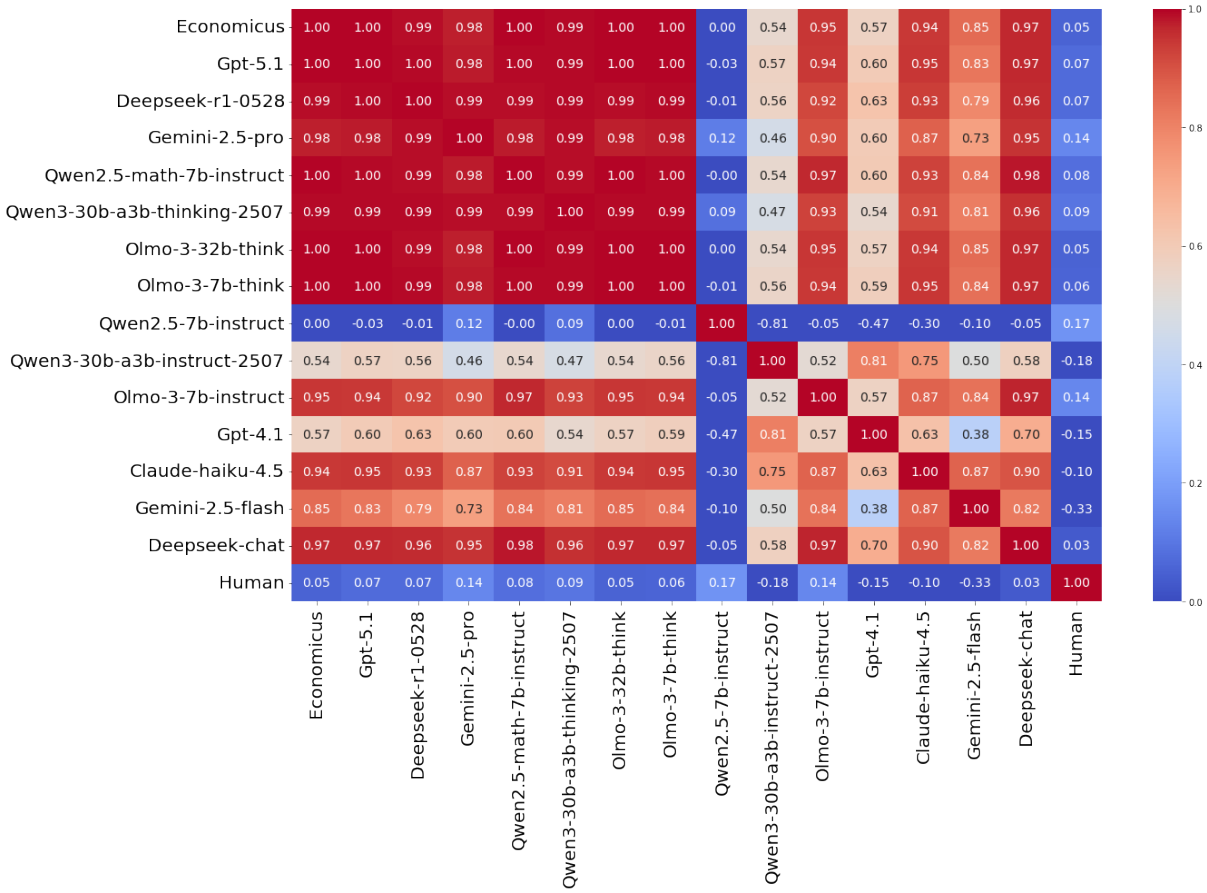

Figure 8: Correlation matrix involving (1) LLMs, (2) *economicus*, and (3) human responses to all questions with explicit prospects.

2, we report the shape parameters (fixing $\beta = 1$).

Model	Model 1 β	Model 2 (Shape-Only) σ	λ	γ
DeepSeek-R1	1000	391	3.19	1.35
Gemini-Pro	745	0.74	1000	0.86
GPT-5.1	1000	338	4.13	1.34
Claude Haiku	14	277	1.26	1.21
DeepSeek-Chat	95	294	1.41	1.49
Gemini-Flash	18	14	2.03	1.39
GPT-4.1	0.01	0.81	70	0.69
Human	0.01	0.63	4.42	0.36

Table 11: Restricted prospect theory parameters of black-box models for decisions with explicit prospects.

Model	Model 1 β	Model 2 (Shape-Only) σ	λ	γ
DeepSeek-R1	407	1.24	506	1.32
Gemini-Pro	127	1.29	132	1.39
GPT-5.1	1000	1.32	1000	1.35
Claude Haiku	11	0.38	13	12
DeepSeek-Chat	6	0.21	10	25
Gemini-Flash	9	1.56	13	0.81
GPT-4.1	14	0.02	1000	0.15
Human	6	0.47	4.57	2.30

Table 12: Restricted prospect theory parameters of black-box models for decisions with implicit prospects.

C.6.2 Parameter Estimates: Regret Aversion

Table 13 and 14 present the fitted parameters for Model 4, the regret aversion model. Here, λ_{reg} represents decisiveness in the regret framework, while κ and α control the shape of the regret function.

We observe that the regret aversion model does consistently poorly compared to the prospect theory model used in the main analysis. In terms of goodness-of-fit, the regret model yields significantly lower correlations and higher MSE across both prospect representations.

C.7 Standard Prospect Theory Results

As discussed in Appendix B.2.1, we also fit a standard prospect theory parameterization with four parameters: σ (risk preference), λ (loss aversion), γ (probability weighting), and β (decisiveness). While this formulation is widely used, λ and β are not jointly identifiable in pure-loss prospects, as the model can only recover their product (see

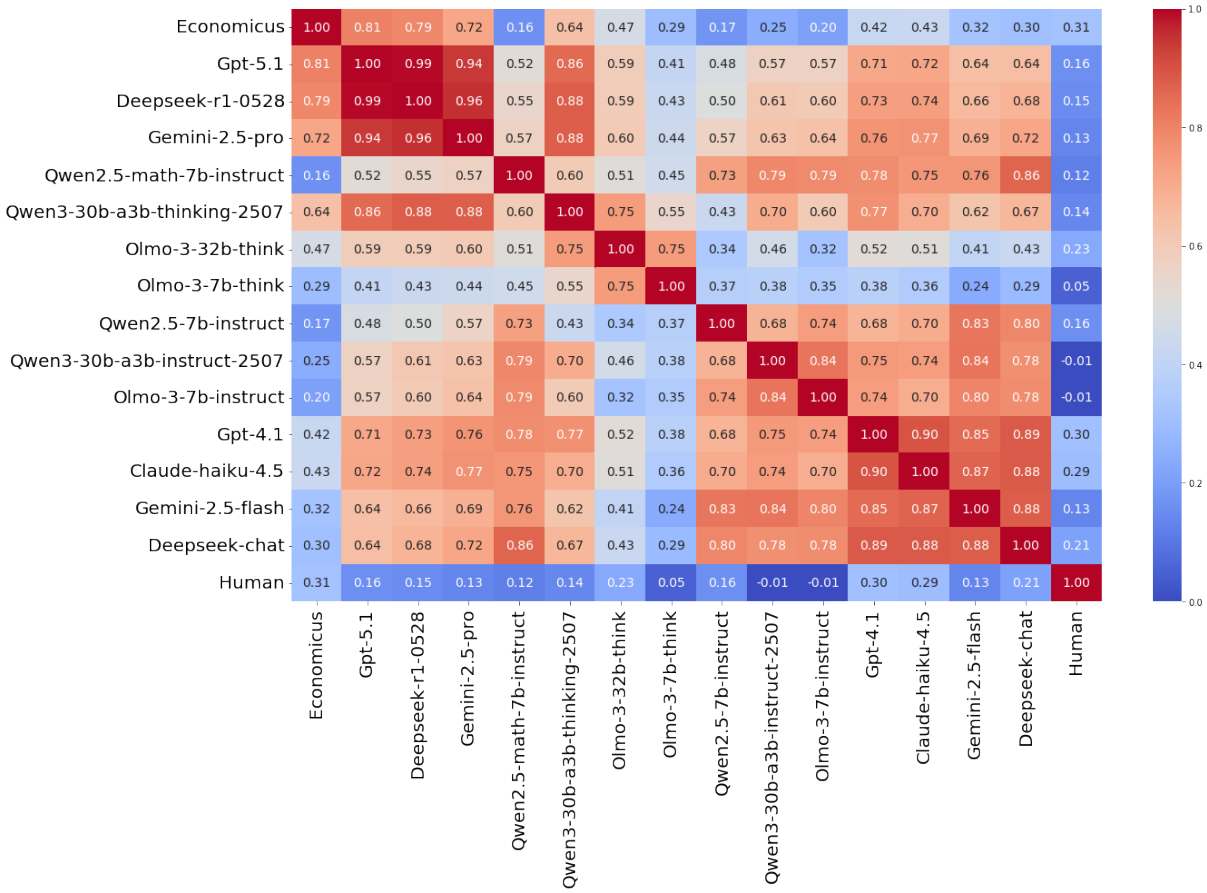

Figure 9: Correlation matrix involving (1) LLMs, (2) *economicus*, and (3) human responses to all questions with implicit prospects.

Model	λ_{reg}	κ	α	Corr	MSE
DeepSeek-R1	0.01	108	0	0.53	0.14
Gemini-Pro	0.01	87	0	0.45	0.15
GPT-5.1	0.01	144	0	0.61	0.15
Claude Haiku	0.01	82	0	0.58	0.07
DeepSeek-Chat	0.01	61	0	0.48	0.08
Gemini-Flash	0.01	127	0.31	0.77	0.03
GPT-4.1	0.01	994	538	-0.26	0.10
Human	0.01	981	81	-0.24	0.01

Table 13: Regret aversion parameters of black-box models for decisions with explicit prospects.

Model	λ_{reg}	κ	α	Corr	MSE
DeepSeek-R1	0.01	51	0	0.30	0.19
Gemini-Pro	0.01	46	0	0.30	0.17
GPT-5.1	0.01	59	0	0.33	0.21
Claude Haiku	0.02	780	181	0.00	0.11
DeepSeek-Chat	0.01	0	2.25	-0.11	0.07
Gemini-Flash	0.81	0	1.32	0.25	0.07
GPT-4.1	0.01	0	2.41	-0.03	0.17
Human	0.06	825	24	0.02	0.03

Table 14: Regret aversion parameters of black-box models for decisions with implicit prospects.

Section B.2.1 for details). We report these results here for completeness; qualitative conclusions are consistent with the dual-beta specification reported in the main text.

C.7.1 Frontier Models

Model	σ	λ	γ	β	Corr	MSE
DeepSeek-R1	0.79	1000	0.86	1000	0.98	0.012
Gemini-Pro	0.82	0.57	0.83	1000	0.95	0.019
GPT-5.1	1.33	1000	1.16	1000	1.00	0.002
Claude Haiku	0.92	3.89	0.83	89	0.72	0.063
DeepSeek-Chat	0.84	0.43	0.84	159	0.75	0.042
Gemini-Flash	0.95	1.58	0.94	20	0.87	0.015
GPT-4.1	0.51	0.01	1.30	150	0.53	0.053
Human	0.04	66	0.13	0.49	0.67	0.004

Table 15: Standard prospect theory parameters for frontier models with explicit prospects.

C.7.2 Open Models

D Human Instructions

As all the participants were recruited via Prolific, they were provided informed consent prior to participation. The consent form described the nature of the study, the types of data collected, and

Figure 10: Consistency and decisiveness heatmap for all models and the human subjects.

Figure 11: Consistency and decisiveness heatmap for all models and the human subjects restricted to the explicit prospect setting.

Figure 12: Consistency and decisiveness heatmap for all models and the human subjects restricted to the implicit prospect setting.

Model	σ	λ	γ	β	Corr	MSE
DeepSeek-R1	1.05	1.20	0.92	1000	0.99	0.005
Gemini-Pro	0.84	1.05	1.24	128	0.95	0.021
GPT-5.1	1.04	1000	0.94	1000	1.00	0.001
Claude Haiku	0.70	0.90	2.10	12	0.66	0.059
DeepSeek-Chat	0.80	0.48	2.51	9	0.59	0.039
Gemini-Flash	1.31	0.91	0.67	17	0.69	0.037
GPT-4.1	1.12	3.42	0.64	18	0.69	0.080
Human	1.03	0.89	0.83	9	0.55	0.022

Table 16: Standard prospect theory parameters for frontier models with implicit prospects.

Model	σ	λ	γ	β	Corr	MSE
Qwen2.5-7B-Instruct	0.58	0.80	3.59	8	0.83	0.017
Qwen2.5-Math-7B	1.11	0.34	1.03	1000	0.99	0.004
Qwen3-30B-Instruct	0.94	0.01	0.86	48	0.57	0.041
Qwen3-30B-Thinking	0.87	1000	0.81	1000	0.99	0.003
Olmo-3-7B-Instruct	0.87	0.20	0.82	220	0.73	0.042
Olmo-3-7B-Think	1.33	1000	1.15	1000	1.00	0.001
Olmo-3-32B-Think	1.33	1000	1.16	1000	1.00	0.000

Table 17: Standard prospect theory parameters for open models with explicit prospects.

how the data would be used and stored. Participants were informed that their responses would be recorded in anonymized form. Once they consented, they were presented with the following instructions prior to completing the task. Then in

Model	σ	λ	γ	β	Corr	MSE
Qwen2.5-7B-Instruct	1.60	0.21	3.17	19	0.58	0.056
Qwen2.5-Math-7B	0.55	0.12	0.88	7	0.54	0.027
Qwen3-30B-Instruct	1.11	0.19	0.81	17	0.50	0.047
Qwen3-30B-Thinking	0.98	1.67	0.85	913	0.95	0.024
Olmo-3-7B-Instruct	1.25	0.27	1.00	4.80	0.44	0.020
Olmo-3-7B-Think	0.82	0.69	0.78	702	1.00	0.001
Olmo-3-32B-Think	0.76	1000	1.31	405	0.95	0.023

Table 18: Standard prospect theory parameters for open models with implicit prospects.

every page, they were presented with a question at a time with two options A,B.

> We invite you to participate in a research study being conducted by investigators from ***.
>
> If you have any questions about the research study itself, please contact: ***. If you have questions, concerns, or complaints about your rights as a research participant, please contact: ***.
>
> Thank you very much for your consideration of this research study.
>
> In each of the following 6 questions, you will be provided with two options

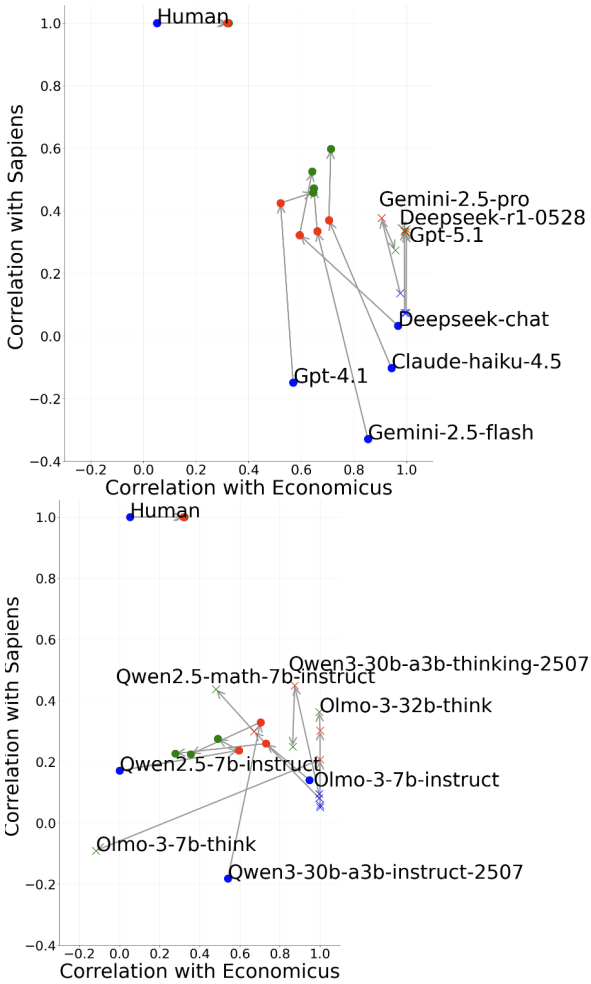

Figure 13: HE representation of the LLMs and the impact of sample size in implicit prospect. Blue: Explicit Prospect; Red: Sample size 100; Green: Sample size 20.

with different payoffs and uncertainties. Please choose the one you prefer.