How Personality Traits Shape LLM Risk-Taking Behaviour

John Hartley[1], Conor Hamill[1], Dale Seddon[1], Devesh Batra[1], Ramin Okhrati[2], Raad Khraishi[1,2]

[1]NatWest AI Research, [2]University College London

Correspondence: john.hartley@natwest.com

Abstract

Large Language Models (LLMs) are increasingly deployed as autonomous agents for simulation and decision-making, necessitating a deeper understanding of their decision-making behaviour under risk. We investigate the relationship between LLMs' personality traits and risk-propensity, applying Cumulative Prospect Theory (CPT) and the Big Five personality framework. We compare the behaviour of several LLMs to human baselines. Our findings show that the majority of the models investigated are risk-neutral rational agents, whilst displaying higher Conscientiousness and Agreeableness traits, coupled with lower Neuroticism. Interventions on Big Five traits, particularly Openness, influence the risk-propensity of several LLMs. Advanced models mirror human personality-risk patterns, suggesting that cognitive biases can be surfaced by optimal prompting. However, their distilled variants show no cognitive bias, suggesting limitations to knowledge transfer processes. Notably, Openness emerges as the most influential factor to risk-propensity, aligning with human baselines. In contrast, less advanced models demonstrate inconsistent generalization of the personality-risk relationship. This research advances our understanding of LLM behaviour under risk and highlights the potential and limitations of personality-based interventions in shaping LLM decision-making.

1 Introduction

Large language models (LLMs) have emerged as powerful autonomous agents demonstrating versatility across diverse domains (Wang et al., 2024) such as software engineering (Cognition.ai, 2024; Xia et al., 2024; Qian et al., 2024, 2023), financial modelling (Lakkaraju et al., 2023; Ding et al., 2024; Yu et al., 2024), and multi-agent simulations of emergent human-like behaviour (Park et al., 2023, 2024; Vallinder and Hughes, 2024; Park et al., 2024). In the financial sector for example, agent-based modelling is already utilized in banking (Hamill et al., 2025), and with the advancement of LLMs, we anticipate an increased application in financial simulations (Li et al., 2024; Zhang et al., 2024; Gao et al., 2024). The success of LLMs can be attributed to their ability to efficiently leverage knowledge from vast text corpora and their proficiency in natural language understanding and generation (Bubeck et al., 2023). These capabilities have led to the development of sophisticated agents with advanced decision-making abilities, incorporating linguistic-based modules for planning, memory, perception, and action (Xi et al., 2023; Park et al., 2023).

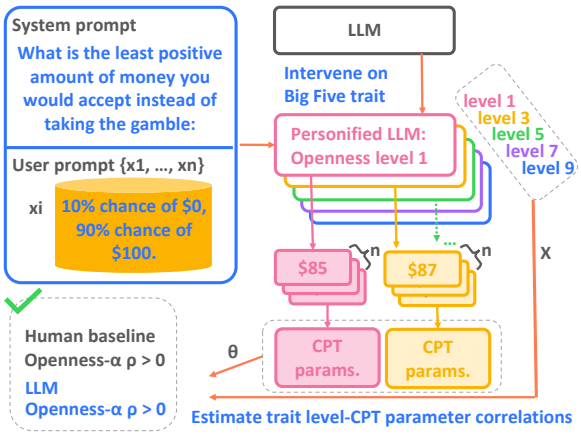


Figure 1: We measure the effect of interventions on the Big Five personality traits (X) to LLMs' CPT parameters (θ). We find analogous relationships between traits and risk-taking in GPT-4o and Claude 3 Sonnet with humans. Parameter α is the risk sensitivity for gains, and ρ is the Spearman's correlation coefficient.

Despite their growing utility and complexity, a comprehensive understanding of LLMs' decision-making processes, in the context of risk, remains an emerging area of research (Binz and Schulz, 2023b; Hagendorff, 2023; Ross et al., 2024; Binz

Findings of the Association for Computational Linguistics: ACL 2025, pages 21068–21092

et al., 2024). Concurrently, LLMs can be prompted to exhibit specific personality traits, such as those defined by the OCEAN model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) (Serapio-García et al., 2023) and can be prompted to role-play with varying personalities (Serapio-García et al., 2023; Jiang et al., 2024). This finding raises intriguing questions about the interplay between personality traits and decision-making processes in LLMs, especially in domains where risk-tolerance and personality characteristics play crucial roles, such as financial applications and real world agents.

While extensive research has been conducted on the connection between human personality and risk behaviour in prospect selection (Nicholson et al., 2005; Soane and Chmiel, 2005; Weller and Thulin, 2012; Boyce et al., 2016; Rustichini et al., 2016; Oehler and Wedlich, 2018; De Bortoli et al., 2019; Joseph and Zhang, 2021; Highhouse et al., 2022), the analogous relationship in LLMs remains unexplored.

We investigate the relationship between personality traits and risky decision-making in LLMs to bridge this gap. We pose the research question: *Do LLMs generalize the relationship between personality traits and risk-based decision-making in a manner analogous to humans?* We explore this question by we applying two well-established frameworks: the Cumulative Prospect Theory (CPT) from behavioural economics (Tversky and Kahneman, 1992) and the Big Five personality traits from psychological testing (Goldberg et al., 1999).

Our approach consists of three main steps.

1. We establish a baseline personality profile for LLMs via a self-reported survey on the IPIP-NEO-300 personality inventory (Goldberg et al., 1999).
2. We create counterfactual scenarios by systematically manipulating the Big Five personality traits in LLMs across various intensity levels.
3. We measure risk-propensity in these counterfactual LLM personas by estimating CPT model parameters based on preferences between certain outcomes and risky gambles. We quantify these causal relationships using Spearman's correlation coefficients and compare them to human baselines (see Figure 1), we ask: *Do the causal links between personality and risk-taking in LLMs mirror those observed in humans?*

Our primary contributions are:

1. We introduce a novel method to estimate the certainty equivalents of LLMs.
2. We establish that LLMs are generally risk-neutral rational agents when evaluating risky prospects. Our approach significantly outperforms existing methods (Ross et al., 2024) in terms of stability, particularly for evaluations involving potential losses.
3. Our analysis shows that LLMs exhibit notably higher levels of Conscientiousness and Agreeableness, along with lower levels of Neuroticism, compared to a sample of human personality measurements.
4. Our research illustrates that targeted interventions on the Big Five personality traits of GPT-4o and Claude 3 Sonnet lead to irrational risk-averse and risk-seeking behaviors in risk assessments, which align risk-propensity patterns observed in human subjects. Whereas we find that Claude 3 Haiku, Gemini 1.5 Pro and Gemini 1.5 Flash generate anti-patterns to human subjects.
5. We establish that many of the smaller model variants do not preserve cognitive biases found their larger variants.
6. We establish that Openness is the most significant personality trait for predicting risk-propensity in GPT-4o, aligning with findings from human behavioural studies (Rustichini et al., 2016).
7. We show that GPT-4 Turbo does generalise the personality-risk relationship. The legacy model yields a monotonic personality-risk relationship for personality markers with variable qualifiers but not across antonymic personality markers.

The paper is structured as follows: Section 2 reviews relevant literature on personality prompting and CPT parameters in LLMs. Section 3 outlines our approach to estimating certainty equivalents of personified LLMs. Section 4 presents our findings on the personality traits and CPT parameters of multiple LLMs, including for personified versions and comparisons with baselines of human behaviour. Section 5 concludes with implications of our research.

2 Related Work

We review related works in psychometric assessment of personality, risk-propensity analysis through prospect theory (PT), and personality modulation in LLMs. We present a comprehensive overview in Appendix A.

The Big Five model factors personality into five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) (McCrae and John, 1992; Costa and McCrae, 1992). Human alignment with these traits is measured through surveying associations with specific personality facets of these traits (Goldberg et al., 1999). Research has shown correlations between the Big Five traits and risk-propensity. For example, **Openness** consistently predicts increased and reduced risk-seeking for gains and losses respectively (Nicholson et al., 2005; Rustichini et al., 2016; Highhouse et al., 2022). **Extraversion** positively correlates with risk-seeking (Nicholson et al., 2005; Oehler and Wedlich, 2018). **Neuroticism** inversely relates to risk-seeking (Nicholson et al., 2005; Soane and Chmiel, 2005). **Agreeableness** generally correlates with lower risk-seeking tendencies (Nicholson et al., 2005; Joseph and Zhang, 2021). **Conscientiousness** inversely relates to risk-seeking (Nicholson et al., 2005; Weller and Thulin, 2012; Boyce et al., 2016).

Machine Psychology is the study of emergent cognitive behaviors in LLMs (Hagendorff, 2023). Recent studies have shown that cognitive biases observed in earlier models have diminished in the latest generation of LLMs (Hagendorff et al., 2023; Chen et al., 2023). Salewski et al. (2024) demonstrated improved task performance in LLMs when given personas, mimicking human-like developmental exploration and expert reasoning. Procedural prompting must be used to prevent text generations using shortcut learning or memorization (Carlini et al., 2021).

Prospect Theory (PT) (Kahneman and Tversky, 1979) describes human decision-making under risk, challenging Expected Utility Theory. Cumulative Prospect Theory (Tversky and Kahneman, 1992) introduced a power-law model to quantify irrational decision-making behaviour. Recent research has investigated PT in LLMs. Binz and Schulz (2023b) found human-like cognitive biases in GPT-3, while Ross et al. (2024) observed reduced biases in more recent models like GPT-4. The latter also showed that risk-propensity is sensitive to qualitative personas.

Big Five traits have been induced in LLMs with demonstrated generalization to downstream tasks (Jiang et al., 2024; Serapio-García et al., 2023). Research has expanded beyond the Big Five to examine toxic traits (Li et al., 2022) and anxiety (Coda-Forno et al., 2023) in language models. While Sühr et al. (2023) identified inconsistencies in earlier models' personality assessments, more recent models like GPT-4 Turbo exhibited greater trait consistency. Huang et al. (2024) demonstrated internal-consistency reliability in LLMs' Big Five scores across 2,500 variations in wording, language, format, and item order. These scores showed significantly lower variance than human norms. This low-variance pattern was reproduced across multiple architectures (GPT-3.5, GPT-4, Gemini-Pro, and Llama-3), strongly suggesting the effect is architecture-agnostic. When deliberately manipulated through extreme trait instructions or character role-playing, personality inventories shifted predictably, confirming the scale's sensitivity to meaningful personality changes while maintaining robustness against irrelevant prompt variations.

We build on these works to demonstrate how intervening on the Big Five personality traits of LLMs affects their risk-based decision-making within the CPT framework. Additionally, we examine whether the relationships between personality traits and risk-taking in LLMs align with those observed in humans.

3 Methodology

3.1 Estimating CPT parameters

Prospect theory describes how humans make decisions under risk. Agents select prospects with the highest utility. The utility of a prospect is calculated by combining subjective probabilities and values of outcomes as follows:

$$u(P) = \sum_{i=1}^{n} w(p_i)v(x_i) \tag{1}$$

where x_i represents an outcome with an associated probability p_i. The set P, defined as $\{x_i, p_i\}_{i=1}^{n}$ encompasses all the possible outcomes and their corresponding probabilities within a prospect. The function $w(p_i)$ serves as the probability weighting function, transforming objective probabilities into

decision weights. Complimenting this, $v(x_i)$ acts as the value function, assigning a subjective value to each outcome.

The value function and the weighting function are parameterised by the CPT parameters $(\alpha, \beta, \lambda, \gamma)$ as follows:

$$v(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \\ -\lambda(-x)^{\beta} & \text{if } x < 0 \end{cases} \tag{2}$$

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}} \tag{3}$$

The CPT parameters describe the risk-propensity of an agent as follows:

- Gain sensitivity (α) - Diminishing sensitivity to gains (risk-averse) for $0 < \alpha < 1$.
- Loss sensitivity (β) - Diminishing sensitivity to losses (risk-seeking) for $0 < \beta < 1$.
- Loss aversion (λ) - Losses loom larger than equivalent gains for $\lambda > 1$.
- Probability sensitivity (γ) - Individuals overweight small probabilities and underweight large probabilities for $0 < \gamma < 1$. For gains and losses the probability sensitivity is denoted by ϕ_+, ϕ_- respectively.

We estimate an agent's CPT parameters by analysing its preferences for certain outcomes to risky prospects. For instance, consider a choice between a risky prospect. E.g. *10% chance of $0, 90% chance of $100 and a certain outcome*. A risk-averse agent prefers a certain outcome below the risky prospect's expected value, while a risk-seeking agent favours one about it. The certain outcome at which the agent is indifferent is called the certainty equivalent.

Our approach to determining CPT parameters is as follows. First, we ask the agent to return its certainty equivalent to each prospect in the dataset (see Section 3.2 for dataset details). Second, we perform non-linear regression on these observed certainty equivalents.

We solve for the CPT parameters that minimize the regression problem in equation (4). The optimization procedure is initialized using median human CPT parameters from Tversky and Kahneman (1992): $(\alpha, \beta, \lambda, \phi_+, \phi_-) = (0.88, 0.88, 2.25, 0.61, 0.69)$. To ensure u accurately represents human behaviour, we constrain all parameters to non-negative values. We employ Nelder-Mead optimization (Nelder and Mead, 1965) from SciPy (Virtanen et al., 2020) for the regression.

$$\theta^* = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left(c_i - v^{-1}[u(P_i|\theta)]\right)^2 \tag{4}$$

$$\text{subject to} \quad \theta_j > 0, \quad \forall j \in \{1, \ldots, m\}$$

where θ are the agent's CPT parameters, c_i is the agent's certainty equivalent to prospect P_i, $u(P_i|\theta)$ is the utility of a prospect given CPT parameters θ, v^{-1} is the inverse function of v, n is the number of risky prospects presented to the agent, and m is the number of CPT model parameters.

3.2 Estimating the CPT parameters of LLMs

Ross et al. (2024) adapted a method from Tversky and Kahneman (1992) to estimate an LLM's certainty equivalents to risky prospects. The two-round process presents the LLM with a prospect and seven logarithmically spaced certain outcomes. The LLM accepts or rejects these in the first round, followed by a refinement using linearly spaced outcomes in the second round. The final certainty equivalent is the midpoint of the accepted/rejected outcome. This process is repeated for all prospects, with CPT parameters estimated using non-linear regression.

There are several issues in this prior approach: misinterpretation of certainty equivalents for prospects with negative expected values, imprecise estimates for large outcomes due to log-spacing, and instabilities when intervening on personality traits, leading to uniform acceptance or rejection of all certainty equivalents.

Our method uses LLM-reported certainty equivalents (see Figure 4 and Figure 5 in Appendix B.3 for the system and user prompts respectively). Our prompt distinguishes between positive and negative expected value prospects by requesting explicitly the *least positive* or *most negative* certain cash amounts respectively. We infer the certainty equivalents over 15 runs with a unique seed per run (see Appendix B.2 for supplementary experimental details). CPT parameters are estimated using equation (4). Figure 9 and Figure 10 in Appendix D.1 show comparisons of certainty equivalents for our work and Ross et al. (2024) respectively. Our approach reduces the number of outliers for prospects with negative expected values, likely due to its simplicity.

We estimate the agent's CPT parameters using 56 prospects from the choices13k dataset (Peterson et al., 2021). These prospects, denoted D, are listed in Table 8 in Appendix C. Our selection of D is motivated by two goals:

1. *Precise loss aversion estimates*: We sample mixed prospects (containing gains and losses) from choices13k, required for estimating the loss aversion parameter λ. The 56 prospects used by Ross et al. (2024), selected from Tversky and Kahneman (1992), are not mixed. This accounts for their imprecise loss aversion parameter estimates, as the optimization converges independently of λ.
2. *Mitigation of prediction biases*: The choices13k data does not contain human-estimated certainty equivalents, reducing the risk of LLMs recalling pre-existing human judgments (Carlini et al., 2019). Whereas the 56 prospects listed in Tversky and Kahneman (1992) do contain certainty equivalents and are likely to be in-sample to training datasets.

3.3 Personality interventions on risk-taking

This section outlines our approach to measure personality traits in LLMs, intervene on LLM personality traits and assess interventional effects on risk-propensity.

Studies show LLMs' personalities can be measured through psychometric testing (Karra et al., 2022) and induced using in-context learning (Serapio-García et al., 2023; Jiang et al., 2024). These induced traits generalize beyond testing, correlating with personality levels in generated content like social media posts (Serapio-García et al., 2023; Jiang et al., 2024). We compare the personality traits of multiple LLMs with a human sample using the IPIP-NEO-300 inventory (Goldberg et al., 1999; Johnson, 2020). This 300-question assessment measures Big Five traits on a 1–5 scale. Scores are calculated by summing responses for related facets (see Table 6 in Appendix B.4). We compare scores to a sample of UK and Irish citizens aged 30+ (Johnson, 2020).

We examine the personality-risk relationship by independently manipulating Big Five traits and measuring CPT parameters. We modify the system prompt with per trait-based persona instructions, as per Serapio-García et al. (2023). Figure 6 in Appendix B.3 illustrates our modified system prompt structure. It combines bipolar adjective markers (such as intelligent-unintelligent), shown effective for Big Five trait characterization by Goldberg et al. (1999), with a 1–9 Likert scale (see Figure 7 in Appendix B.4) to regulate marker intensity. This approach, adapted from Serapio-García et al. (2023), allows precise independent manipulation of personality traits in LLMs. Markers represent specific trait facets, categorized as low-level (1–3), neutral (4), or high-level (5–9). Table 7 in Appendix B.4 shows a comprehensive list of these markers.

We employ counterfactual analysis to examine the relationship between personality traits and CPT parameters across various LLMs. By systematically manipulating individual trait levels while holding others constant, we create "what-if" scenarios to isolate each trait's impact on risk-taking behaviour. We then calculate Spearman's correlation coefficients between these manipulated trait levels and the resulting CPT parameters, examining their polarity and statistical significance. This counterfactual approach allows us to quantify the causal influence of each trait on risk-taking behaviour, identify potential relationships, and compare these trait-risk associations in LLMs to established human behavioural patterns.

4 Experiments and results

We present our experiments and results in three parts. We begin by examining the CPT parameters and personality traits of LLMs. Next, we explore risk-taking behaviour in LLMs with interventions on their personality traits. Finally, we investigate how Openness emerges as the most influential trait in risk-taking.

4.1 LLM personality and risk-taking

In this section, we investigate the risk-propensity of LLMs (Figure 2a), suggest explanations based on the their personality traits (Figure 2b), and compare against baselines of human behaviour. Our findings reveal both similarities and differences, contributing to the understanding of LLMs' decision-making under risk.

4.1.1 Risk-neutral rational agents

Our analysis begins with an examination of risk-propensity patterns across various LLMs. Figure 2a illustrates that GPT-4o, Claude 3 Sonnet, Gemini 1.5 Pro, and Gemini 1.5 Flash demonstrate median CPT parameters approximating unity within their

confidence intervals (see Table 10 in Appendix D.3 for precise numerical values). Parameter values equal to one indicate that these models behave as risk-neutral rational agents in prospect selection.

In contrast, smaller model variants, GPT-4o mini and Claude 3 Haiku, exhibit less stable risk-taking behaviour, evidenced by wider confidence intervals for their CPT estimates (Figure 2a). These models show poorer alignment with risk-neutral rational agents, in terms of loss aversion ($\lambda = 0.74$, $\lambda = 1.19$ respectively) and probability distortion for losses ($\phi_- = 0.93$, $\phi_- = 0.86$ respectively).

4.1.2 Personality traits of LLMs

Figure 2b shows that LLMs generally exhibit higher Openness, Conscientiousness and Agreeableness, coupled with lower Neuroticism, compared to the human sample from Johnson (2020). These trait profiles typically correlate with reduced risk-taking behaviour, aligning with the observed risk-neutral patterns in most models we examined.

For example, the mean Openness score observed for GPT-4o is significantly greater than the mean human score from Johnson (2020) (one-tailed t-test p-value < 0.05). Moreover, in Figure 2b we observe statistically significant greater levels of Openness in all five other models (GPT-4o-mini, Claude 3 Sonnet, Claude 3 Haiku, Gemini 1.5 Pro, Gemini 1.5 Flash) in comparison to the human score. This result strongly suggests that frontier LLMs have greater Openness scores than humans.

We hypothesize that increased Openness is attributed to reward model (RM) training on human feedback data. For example, Gemini Team et al. (2023) note that RMs, which weight training examples preferred by humans in alignment processes, are optimised using human feedback data where human annotators are asked to favour responses exhibiting creativity — a key facet of Openness.

4.1.3 GPT-4o personality traits and risk behavior

GPT-4o demonstrates higher risk-taking for gains and lower risk-taking for losses (Figure 2a). Highhouse et al. (2022) reported that Openness is the most influential personality trait in risk-propensity in humans, accounting for 22% of the variance in risk-propensity, correlating positively with risk-taking for potential gains and negatively for potential losses (Rustichini et al., 2016). This suggests parallels in decision-making processes between GPT-4o and humans.

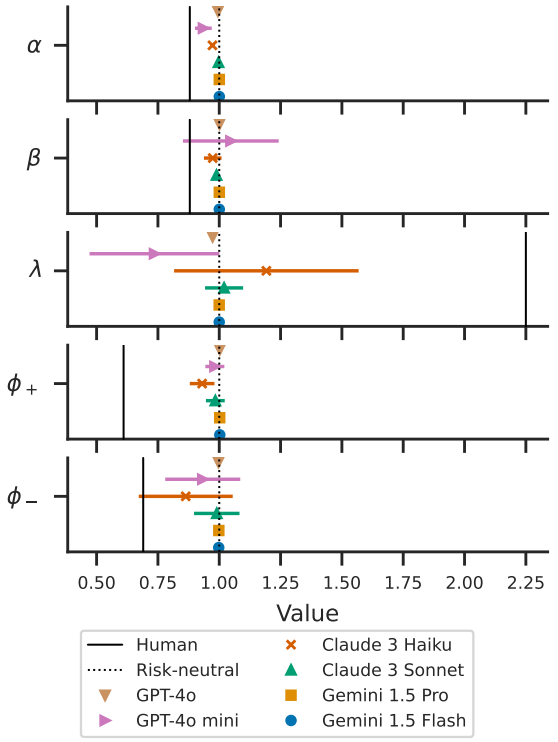
While Openness plays a primary role, other personality traits: Conscientiousness, Agreeableness, Extraversion, and Neuroticism—also influence risk-propensity, albeit to a lesser extent. Previous studies by Nicholson et al. (2005) and Joseph and Zhang (2021) indicate that higher Conscientiousness and Agreeableness is associated with lower risk-seeking in humans.

Figure 2b reveals GPT-4o exhibits significantly higher Conscientiousness and Agreeableness scores compared to human baselines. Figure 2a demonstrates GPT-4o's lower risk-seeking for losses (β) relative to human averages. Additionally, GPT-4o displays risk-averse behaviour in an absolute sense for gains (α). These patterns indicate that elevated Conscientiousness and Agreeableness scores correspond with reduced risk-seeking behaviors in GPT-4o. Such trait-behaviour correlations mirror established relationships documented in human psychological literature (Nicholson et al., 2005; Joseph and Zhang, 2021).

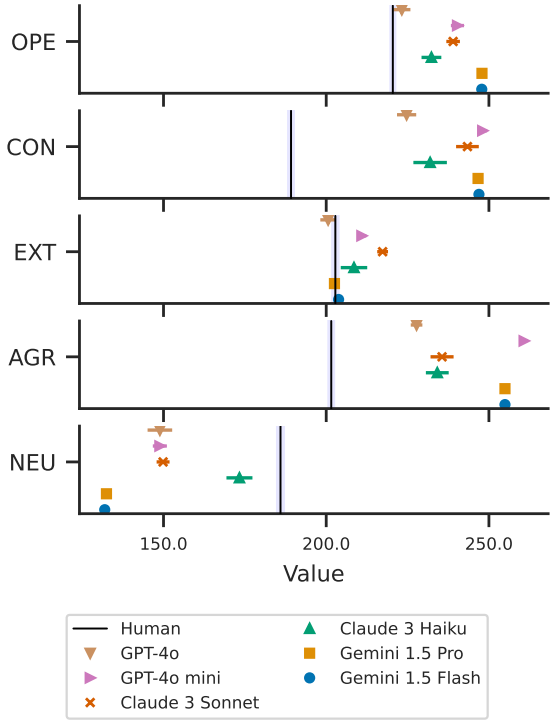
Nicholson et al. (2005); Soane and Chmiel (2005); Oehler and Wedlich (2018) showed that lower Neuroticism is associated with higher risk-taking in humans. Our results indicate that GPT-4o exhibits a significantly lower level of Neuroticism than the human sample (Figure 2b), and demonstrates a lower level of risk-aversion for gains than the humans on average, as evidenced by $\alpha = 0.99$ in Figure 2a, suggesting that the Neuroticism-risk relationship is mirrored by GPT-4o for gains.

4.2 Risk-taking in personified LLMs

This experiment explores the relationship between Openness and risk-propensity, comparing it to human patterns. We independently manipulate the Openness trait across multiple intensity levels (1, 3, 7, and 9) and estimate the CPT parameters using gambles from dataset D over multiple models (GPT-4o, GPT-4o-mini, Claude 3 Sonnet, Claude 3 Haiku, Gemini 1.5 Pro, and Gemini 1.5 Flash). Table 11 in Appendix E shows that the spillover to other traits resulting from the intervention is minimal. Table 13 in Appendix E shows that the effect size monotonically increases with the intervention level. We calculate Spearman's correlation coefficients to assess the monotonic relationship between Openness and each CPT parameter. Table 1 presents our findings.



(a) CPT parameter estimates: LLMs (15 runs with random seeds) and humans (Tversky and Kahneman, 1992). Markers show median values and error bars show bootstrapped 95% confidence intervals.



(b) Big Five personality trait scores (IPIP-NEO-300): LLMs (15 runs with random seeds) and humans (Johnson, 2020) Markers represent means with error bars and shaded regions indicating 95% confidence intervals.

Figure 2: Estimates of CPT parameters and personality scores for LLMs and humans.

Model	α	β	λ
GPT-4o	0.52^{***}	0.44^{**}	-0.30^{*}
GPT-4o mini	0.06	-0.12	-0.10
GPT-4 Turbo	-0.17	-0.01	-0.24
Claude 3 Sonnet	0.41^{***}	0.46^{***}	-0.15
Claude 3 Haiku	0.22	-0.34^{***}	0.47^{***}
Gemini 1.5 Pro	-0.33^{**}	-0.19	-0.21
Gemini 1.5 Flash	-0.51^{***}	-0.32^{**}	-0.06

Table 1: Spearman's correlation coefficients between the level of a personality trait intervention and the model's estimated CPT parameter values over non-mixed prospects in dataset D for multiple LLMs. Parameters α, β, λ are estimated from certainty equivalents using equation (4). The significance of coefficients is determined using t-statistics. Results marked with */**/*** have statistical significance at $\alpha = 0.05/0.025/0.001$ level respectively.

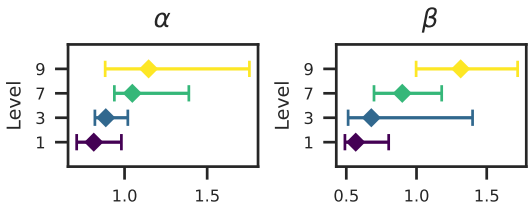
4.2.1 GPT-4o and Claude Sonnet 2 show cognitive biases in risk-taking

Our analysis, presented in Table 1, shows significant correlations between Openness and risk-propensity in large language models (LLMs). GPT-4o and Claude 3 Sonnet exhibit positive correlations with risk-seeking for gains ($\rho = 0.52$, $\rho = 0.41$ respectively) and risk-aversion for losses ($\rho = 0.44$, $\rho = 0.46$ respectively), mirroring human behaviour. These findings contradict recent suggestions by Hagendorff et al. (2023) and Chen et al. (2023) that cognitive biases have diminished in current LLMs. While the baseline behaviour (without personality interventions) of these models is relatively risk-neutral (Figure 2a), our personality interventions induce measurable cognitive biases which align with biases in humans.

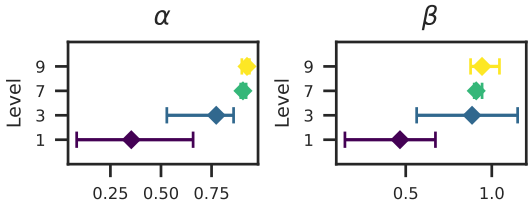
4.2.2 Smaller model variants erase cognitive biases

Interestingly, Table 1 shows that smaller model variants (GPT-4o mini and Claude 3 Haiku) do not replicate this relationship. GPT-4o mini shows no significant correlation, while Claude 3 Haiku demonstrates an inverse correlation between Openness and risk-aversion for losses. This pattern extends to Gemini 1.5 Pro and its smaller variant, both displaying inverse relationships compared to human behaviour, with the effect more pronounced in the smaller model.

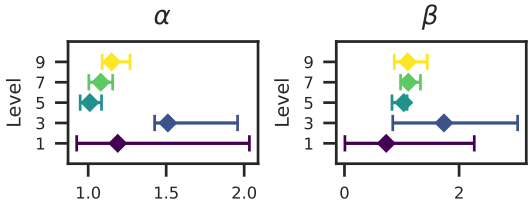
The observed reduction in cognitive bias from larger to smaller models within each model family suggests that knowledge transfer processes from larger to smaller model may not preserve these



(a) Claude 3 Sonnet



(b) GPT-4o



(c) GPT-4 Turbo

Figure 3: CPT parameter estimates with interventions on the Openness level over 15 runs for (a) Claude 3 Sonnet, (b) GPT-4o and (c) GPT-4 Turbo. Markers show median values, and error bars show bootstrapped 95% confidence intervals.

biases. This finding has important implications for model compression techniques, particularly the transfer of latent human-like behaviors in LLMs.

4.2.3 Global and local variations in personality-risk-propensity relationships

Figure 3a and Figure 3b show variations in α and β for Openness interventions for Claude 3 Sonnet and GPT-4o respectively. The results for Claude 3 Sonnet clearly show risk-aversion for low Openness levels and risk-seeking for high levels, and the inverse for β. However, although the correlations between Openness level and risk-sensitivity in GPT-4o are consistent with those observed in human subjects. Our results show that we are unable to produce risk-seeking behaviours for gains or risk-averse behaviours for losses in an absolute sense. This limitation suggests that personality prompting alone is insufficient to elicit these specific risk behaviours in GPT-4o. We emphasize

Trait	α	β	λ
Openness	0.52^{***}	0.44^{**}	-0.30^{*}
Conscientiousness	0.0	-0.06	0.05
Extraversion	0.21	0.15	-0.08
Agreeableness	0.11	0.27	-0.44^{**}
Neuroticism	-0.03	-0.06	-0.04

Table 2: Spearman's ρ for personality traits and CPT parameters in dataset D for GPT-4o. Significance: $^{*}/^{**}/^{***}$ at $\alpha = 0.05/0.025/0.001$ (t-statistics).

the importance of this result for researchers designing personality profiles for agents in multi-agent simulations, such as those proposed by Park et al. (2023), where realistic decision making is required.

We also analysed Openness interventions in GPT-4 Turbo. Figure 3c shows α and β variations across Openness levels. The personality-risk relationship aligns with human behaviour for only trait levels $[1, 3]$ and $[5, 7, 9]$, possibly due to common personality markers in these subgroups (e.g., *Socially Conservative* vs *Socially Progressive*, see Table 7). Unlike GPT-4o's global monotonicity (Figure 3b), GPT-4 Turbo's risk-propensity is not globally monotonic. This suggests that GPT-4 Turbo shows a localized mapping, correlating with human baselines, but only for variations in qualifiers applied to specific markers. This finding demonstrates enhanced cognitive modelling capabilities in more recent GPT models.

4.3 Openness is primary driver of risk-propensity in GPT-4o and humans

Recent studies in human psychology have demonstrated that Openness accounts for 22% of the variance in risk-propensity (Highhouse et al., 2022). We extend this concept to LLMs, specifically examining the influence of personality traits on the risk-propensity of GPT-4o.

To investigate this, we repeat the Openness intervention on the personality traits Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Our findings are presented in Table 2.

Our analysis shows that Openness is the only personality trait significantly correlated with the risk-sensitivity parameters α and β ($\rho = 0.52$ and $\rho = 0.44$, respectively) in GPT-4o. This finding aligns with Highhouse et al. (2022), who identified Openness as the primary personality trait associated with risk propensity in humans. We found no statistically significant correlations be-

tween the other personality traits (Conscientiousness, Extraversion, Agreeableness, and Neuroticism) and the risk sensitivity parameters. This suggests that interventions targeting these traits are unlikely to significantly affect risk propensity in GPT-4o. Notably, this behaviour diverges from human patterns observed in previous studies (Nicholson et al., 2005; Soane and Chmiel, 2005; Weller and Thulin, 2012; Boyce et al., 2016; Oehler and Wedlich, 2018; Joseph and Zhang, 2021). The discrepancy indicates that fine-tuning or more advanced intervention strategies are required to align GPT-4o with human cognitive biases.

5 Conclusion

This study examines the risk-propensity behaviour of LLMs in relation to personality traits. Advanced LLMs predominantly exhibit risk-neutral rational agent characteristics. However, interventions on the Openness trait induce risk-propensity patterns analogous to human behaviour in models such as GPT-4o and Claude 3 Sonnet. Notably, Openness emerges as the primary determinant of risk-propensity in GPT-4o, aligning with human psychological studies. Smaller model variants and legacy models demonstrate inconsistent personality-risk relationships, suggesting potential limitations in knowledge distillation processes. These findings have significant implications for the development of AI systems with human-like decision-making capabilities, particularly in multi-agent simulations and financial modelling.

Limitations

Our research examines how interventions on personality traits affect the risk-propensity of multiple LLMs. Whilst we analysed multiple models, we analysed only two frontier models: GPT-4o and GPT-4o mini. Our analysis could be extended to newer reasoning models that have been recently made available such as DeepSeek and o1-mini.

We assume that the frameworks of CPT and the Big Five personality traits are realistic representations of personality and risk-taking in humans.

Our study assumes the effectiveness of personality interventions across the models we tested. Due to computational constraints and time constraints, we only verified this for a single trait (Openness) in one model (GPT-4o).

To mitigate prediction biases from in-sample data memorization, we selected data points from the choices13k dataset which does not contain human certainty equivalents. However, we cannot guarantee the absence of certainty equivalents for these prospects in public code repositories or other online resources.

We assume that the personality trait score sample of human responses is representative of the population of the UK citizens over 30 (Johnson, 2020). This was the only publicly available data source for the responses to the IPIP-NEO-300 personality inventory available to us. We report the standard deviations of the scores on this data to quantify uncertainty.

We did not include a random baseline condition where heuristic values or trait prompts are randomly assigned. While such a baseline could potentially provide additional insights, the variability in the magnitude of induced traits between different personality dimensions and across intervention levels would likely introduce biases, disproportionately affecting some traits. This consideration led us to focus on per-trait correlations with CPT parameters rather than employing linear modelling approaches that might be more sensitive to such biases. Future work could explore more sophisticated baseline designs that account for these differential effects.

Ethical consideration

Our research investigates the relationship between personality traits and risk-taking behaviour in LLMs in comparison to humans. We exclusively use aggregated descriptive statistics from existing literature, specifically mean and variance of IPIP-NEO-300 personality trait scores made publicly available by Center for Open Science (Johnson, 2020). We do not access or experiment with individual-identifying features. While acknowledging potential demographic biases in these statistics, we believe our research poses minimal risks as we do not employ these data for downstream tasks that could negatively impact any group. Further work should explore demographic biases present in this dataset to understand effects on relationships between personality and risk. Our motivation is to draw comparisons between LLMs and human populations, contributing to understanding AI systems' behavioural patterns without compromising privacy or perpetuating harmful biases.

Acknowledgements

We wish to express appreciation to Greig Cowan, Graham Smith, and Zachery Anderson of NatWest Group for the time and support needed to develop this research paper.

References

Amazon Web Services. 2025. Amazon bedrock.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Michael C Ashton and Kibeom Lee. 2009. The hexaco–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. 2024. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Marcel Binz and Eric Schulz. 2023a. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

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

Christopher J Boyce, Alex M Wood, and Eamonn Ferguson. 2016. Individual differences in loss aversion: Conscientiousness predicts how life satisfaction responds to losses versus gains in income. *Personality and Social Psychology Bulletin*, 42(4):471–484.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Center for Open Science. 2025. Terms of use.

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*, 120(51):e2316205120.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Cognition.ai. 2024. Introducing devin.

Paul T Costa and Robert R McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Daiane De Bortoli, Newton da Costa Jr, Marco Goulart, and Jéssica Campara. 2019. Personality traits and investor profile analysis: A behavioral finance study. *PloS one*, 14(3):e0214062.

John M Digman. 1989. Five robust trait dimensions: Development, stability, and utility. *Journal of personality*, 57(2):195–214.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

Adrian Furnham, Steven C Richards, and Delroy L Paulhus. 2013. The dark triad of personality: A 10 year review. *Social and personality psychology compass*, 7(3):199–216.

Shen Gao, Yuntao Wen, Minghang Zhu, Jianing Wei, Yuhan Cheng, Qunzi Zhang, and Shuo Shang. 2024. Simulating financial market via large language model based agents. *arXiv preprint arXiv:2406.19966*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

GitLab. 2025. Gitlab duo.

Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

Google Cloud. 2025. Vertex ai.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Conor Brian Hamill, Raad Khraishi, Simona Gherghel, Jerrard Lawrence, Salvatore Mercuri, Ramin Okhrati, and Greig Alan Cowan. 2025. Agent-based modelling of credit card promotions. *International Journal of Bank Marketing*.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Scott Highhouse, Yi Wang, and Don C Zhang. 2022. Is risk propensity unique from the big five factors of personality? a meta-analytic investigation. *Journal of Research in Personality*, 98:104206.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of The 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

John A Johnson. 2020. Johnson's ipip-neo data repository.

Elizabeth D Joseph and Don C Zhang. 2021. Personality profile of risk-takers. *Journal of Individual Differences*.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

Kausik Lakkaraju, Sara E Jones, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath C Muppasani, and Biplav Srivastava. 2023. Llms for financial advisement: A fairness and efficacy study in personal decision making. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 100–107.

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)*, pages 15523–15536.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Microsoft. 2025. Azure openai service.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.

John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Nigel Nicholson, Emma Soane, Mark Fenton-O'Creevy, and Paul Willman. 2005. Personality and domain-specific risk taking. *Journal of Risk Research*, 8(2):157–176.

Andreas Oehler and Florian Wedlich. 2018. The relationship of extraversion and neuroticism with risk attitude, risk perception, and return expectations. *Journal of Neuroscience, Psychology, and Economics*, 11(2):63.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

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*, pages 1–22.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

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*, 372(6547):1209–1214.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Melissa J Ree, Davina French, Colin MacLeod, and Vance Locke. 2008. Distinguishing cognitive and somatic dimensions of state and trait anxiety: Development and validation of the state-trait inventory for cognitive and somatic anxiety (sticsa). *Behavioural and Cognitive Psychotherapy*, 36(3):313–332.

Jillian Ross, Yoon Kim, and Andrew W Lo. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Aldo Rustichini, Colin G DeYoung, Jon E Anderson, and Stephen V Burks. 2016. Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation. *Journal of Behavioral and Experimental Economics*, 64:122–137.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Gerard Saucier and Lewis R Goldberg. 1996. The language of personality: Lexical perspectives on the five-factor model.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Sakib Shahriar, Brady D Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Applied Sciences*, 14(17):7782.

Emma Soane and Nik Chmiel. 2005. Are risk preferences consistent?: The influence of decision domain and personality. *Personality and Individual Differences*, 38(8):1781–1791.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. 2023. Challenging the validity of personality tests for large language models. *arXiv e-prints*, pages arXiv–2311.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323.

Aron Vallinder and Edward Hughes. 2024. Cultural evolution of cooperation among llm agents. *arXiv preprint arXiv:2412.10270*.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

John Von Neumann and Oskar Morgenstern. 2007. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Joshua A Weller and Erik W Thulin. 2012. Do honest people take fewer risks? personality correlates of risk-taking to achieve gains and avoid losses in hexaco space. *Personality and individual differences*, 53(7):923–926.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2024. Agentless: Demystifying llm-based software engineering agents. *arXiv preprint arXiv:2407.01489*.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, pages 595–597.

Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. 2024. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*.

A Literature review

In this section we provide a detailed review of related works across machine learning, psychology, and economics in three key areas: psychometric assessment of personality in humans and LLMs, the analysis of risk-propensity through the lens of prospect theory (PT), and the implementation of in-context learning interventions aimed at modulating the personality of LLMs.

A.1 Personality and risk-propensity in humans

Researchers have demonstrated that human personalities can be explained by several independent factors. A widely accepted model is the Big Five, which categorizes personalities into five traits, often referred to by the acronym OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) (McCrae and Costa, 1987; Digman, 1989; McCrae and John, 1992; Costa and McCrae, 1992; Saucier and Goldberg, 1996). An individual's alignment with these factors is measured through their responses to questions associated with specific personality facets for each trait (Costa and McCrae, 1992; Goldberg et al., 1999). For instance, the IPIP-NEO inventory (Goldberg et al., 1999), extensively used in this work, contains several facets for each trait. Taking Conscientiousness as an example, its first four facets are: *Self-efficacy*, *Orderliness*, *Dutifulness*, and *Achievement-Striving* (an exhaustive list of personality traits and their associated personality facets is given in Table 6.).

The following section examines how the Big Five personality traits correlate with risk-propensity in human subjects:

Openness to Experience: The personality trait of Openness is a consistent predictor of risk-taking behaviors in humans. Research studies have shown its significant correlation with increased risk-propensity across various scenarios. For instance, Nicholson et al. (2005) identify Openness as a strong predictor of higher risk-propensity. Rustichini et al. (2016) further delineate that Openness is positively associated with risk-taking in contexts where gains are involved, but negatively associated in scenarios involving potential losses. A meta-analysis conducted by Highhouse et al. (2022) substantiates these findings, reporting that Openness accounts for 22% of the variance in risk-propensity. This indicates a substantial impact of Openness on individual's risk-taking behaviors. Focusing on investment decisions, De Bortoli et al. (2019) reveal that individuals scoring high on Openness tend to adopt riskier investment strategies.

Extraversion: Multiple studies report a positive correlation between extraversion and risk-propensity. Nicholson et al. (2005) and Oehler and Wedlich (2018) find that individuals high in extraversion tend to engage in more risk-taking behaviours.

Neuroticism: Research indicates an inverse relationship between neuroticism and risk-taking. Nicholson et al. (2005) and Soane and Chmiel (2005) find that higher levels of neuroticism are linked to lower risk-taking. Further, Oehler and Wedlich (2018) add that neurotic individuals exhibit greater risk aversion, particularly in investment contexts.

Agreeableness: Higher agreeableness generally correlates with lower risk-taking tendencies. Studies by Nicholson et al. (2005) and Joseph and Zhang (2021) indicate that agreeable individuals are less inclined toward risk. Additionally, Soane and Chmiel (2005) observe that agreeableness is inversely related to risk-taking.

Conscientiousness: Conscientiousness shows a complex relationship with risk. While Nicholson et al. (2005) suggest lower conscientiousness may correspond with higher risk-taking, Weller and Thulin (2012) specifically associate low conscientiousness with greater risk-taking to achieve gains. Boyce et al. (2016) highlight conscientious individuals' strong reactions to income losses, signifying pronounced loss aversion. In investment contexts, Oehler and Wedlich (2018) note that high conscientiousness aligns with greater risk aversion.

A.2 Machine psychology

Emergent cognitive behaviours in LLMs cannot be studied under the typical train-and-test machine learning paradigm since these behaviours are not explicitly coded for during training (Hagendorff, 2023). Instead, tests from behavioural psychology can be used to reveal characteristic behaviour of black-box LLMs without necessarily understanding how the behaviour was learnt. Authors caution that care should be taken not to generalise the results of self-reported psychometric test results beyond any particular system prompt (Röttger

et al., 2024). Nevertheless, several works have shown that induced personalities generalise to new tasks (Serapio-García et al., 2023; Ross et al., 2024).

Binz and Schulz (2023b) reported the existence of human-like cognitive biases in GPT-3. However, more recent studies by Hagendorff et al. (2023) and Chen et al. (2023) noted that these cognitive biases have disappeared in the latest generation of LLMs (post GPT-3.5). These models act as rational agents even without the addition of chain-of-thought prompting. They have also shown that LLMs answer questions using short-cut learning (Geirhos et al., 2020) or memorisation (Carlini et al., 2019, 2021). It is hypothesised that popular psychometric tests are in the training data, and due to the uniqueness of their data, they are likely memorised (Carlini et al., 2019, 2021). These behaviours can be mitigated by using procedurally generated prompts.

Salewski et al. (2024) demonstrated that LLMs improved task performance when given personas. In a multi-armed bandit task, they show that, LLMs impersonating children of various ages mimicked human-like developmental exploration. In a reasoning task, LLMs portraying domain experts outperformed those assuming non-expert roles.

A.2.1 Prospect theory and LLMs

Kahneman and Tversky (1979) introduced prospect theory, a description of how humans make decisions under risk. PT challenges the assumptions of expected utility theory (EUT) (Von Neumann and Morgenstern, 2007) by focusing on observed human behaviour rather than prescriptive models of rational choice. In Kahneman and Tversky (1979) they observed several irrational behaviours which deviated from EUT. Notably, they found that humans exhibit a distorted perception of probability, overweighting low probabilities and underweighting high probabilities. Furthermore, they demonstrated that individuals tend to be risk-averse when facing potential gains but become risk-seeking when confronted with potential losses. Perhaps most significantly, their research revealed that humans are more sensitive to losses than to equivalent gains, a phenomenon known as loss aversion. cumulative prospect theory (Tversky and Kahneman, 1992) introduced a power-law model to quantify the irrational decision making behaviour observed in Kahneman and Tversky (1979). (The model is explained in detail in Section 3).

More recently, several researchers have investigated decision-making in LLMs with respect to prospect theory. Binz and Schulz (2023b) submitted GPT-3 to a number of vignettes and tasks from the cognitive psychology literature. They found that GPT-3 showed several human-like cognitive biases from PT: certainty effect [1], overweighting effect[2], and the framing effect[3]. In subsequent work, Binz and Schulz (2023a) improved the alignment between LLama 2–65B (Touvron et al., 2023) and human decision-making by training a logistic regression model on LLama 2 embeddings of risky prospects from the choices13k dataset (Peterson et al., 2021). They also achieved increased performance on a hold-out task. However, the limitation of this work is that the model is no longer generative. Concurrently (Ross et al., 2024), measured the CPT parameters of GPT-4o and GPT-4-Turbo and found that cognitive biases in these more recent models are reduced. The authors also show that risk-propensity is sensitive to qualitative personas. However, a quantitative comparison with risk personas in humans cannot be made since these profiles are not based on established personality models. In our work we investigate the relationship between established personality models and risk-propensity in LLMs, and examining the similarities between these relationships in humans.

A.3 Personality prompting

Recent research has extensively explored the induction and measurement of personalities in LLMs. For example, Jiang et al. (2024) employed self-reported psychometric testing to assess LLMs' personalities along the Big Five traits. They developed a method to induce these traits in LLMs and demonstrated that the induced personalities generalised to vignette experiments. Similarly, Serapio-García et al. (2023) reported on LLM personality types using psychometric tests. Their method allowed for independent induction of personalities along each trait, with controllable levels for the intensity of each personality trait. The authors validate their method by measuring the correlation between induced personality trait levels and the LLMs' re-

[1]The certainty effect is the tendency of agents to prefer certain outcomes to risky outcomes.

[2]The overweighting effect is the tendency of agents to overweight the probability of rare outcomes.

[3]The framing effect describes the influence of the presentation of outcomes on decision-making.

sponses to the IPIP-NEO-300 inventory for each trait. Furthermore, they observed that these personalities generalised well to downstream tasks such as generating social media posts.

Several studies have explored personality characteristics beyond the Big Five personality traits. Li et al. (2022) reported that LLMs generally score higher than the human average for toxic personality traits, specifically the Dark Triad (Furnham et al., 2013). The authors successfully implemented interventions on these traits using in-context learning. Miotto et al. (2022) assessed GPT-3's personality using the 60-item Hexaco questionnaire (Ashton and Lee, 2009), revealing multiple personalities at different sampling temperatures. Coda-Forno et al. (2023) found that GPT-3.5 displays higher than human average anxiety on the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) Questionnaire (Ree et al., 2008).

The validity of self-reported personality testing in LLMs has recently been questioned. Sühr et al. (2023) demonstrated that GPT-3.5 and LLama 2 do not show consistency in their evaluation of personality inventories. For example, their factor analysis revealed that responses to the Big-Five-Inventory 2 (BFI-2) do not show a simple structure for the first five factors of variation. However, they found that the component loadings for GPT-4-Turbo are separable for each factor, indicating consistent personality traits across questioning for this more advanced model.

B Additional details

Here we supply supplementary materials to support our arguments in the main text. We include additional details on LLMs used in this work, our experimental setup, prompt templates and the Big Five personality traits, and datasets.

B.1 Models and packages

Table 3 provides summary information of the LLMs used in this work. Table 4 reports a summary of the estimated number of tokens used in text-completion APIs for each our main experiments. In addition, we used the AI assistant GitLab Duo to assist in writing our paper (GitLab, 2025). Specifically, we used it only as a writing assistant to rephrase text in our Tex script, and to generate plotting code. In addition, we made use of the following numerical packages given in Table 5.

Model	Model version
GPT-4o	`gpt-4o (2024-08-06)`
GPT-4o mini	`gpt-4o mini (2024-07-18)`
GPT-4 Turbo	`gpt-4 (turbo-2024-04-09)`
Claude 3 Sonnet	`claude-3-sonnet-20240229-v1.0`
Claude 3 Haiku	`claude-3-haiku-20240307-v1.0`
Gemini 1.5 Pro	`gemini-1.5-pro-002`
Gemini 1.5 Flash	`gemini-1.5-flash-002`

Table 3: Summary information of the LLM used in this work. GPT-4o, GPT-4o mini, and GPT-4 Turbo (Shahriar et al., 2024) are provided by Azure OpenAI (Microsoft, 2025); Claude 3 Sonnet and Claude 3 Haiku (Anthropic, 2024) by AWS Bedrock (Amazon Web Services, 2025); Gemini 1.5 Pro and Gemini 1.5 Flash (Gemini Team et al., 2024) by GCP Vertex AI (Google Cloud, 2025).

B.2 Experimental setup

Our experimental methodology involves 15 runs for each experiment, each utilizing a unique random seed for text generation across all services. For example, the random seed remains consistent within each run of the IPIP-NEO-300 survey, and each set of gambles, changing only between runs. It's worth noting that at the time of this study, random seed control was not available for the Gemini 1.5 Pro and Gemini 1.5 Flash text completion APIs. To introduce variance, we set the temperature parameter to 1 for all experiments.

The number of runs (15) was determined based on available computational resources. While previous studies, such as Ross et al. (2024), have employed 100 runs, our research encompasses multiple LLMs and personality trait levels, necessitating a more constrained approach. Despite this limitation, we quantify the uncertainty in all our results throughout the main text.

Our experiments are run on a workstation with an Intel(R) Xeon(R) E3-1585L CPU, and 16GB RAM. To run prompts we use the text-completion APIs listed in Table 3.

B.3 Prompts

This section presents the prompts used in our study. We demonstrate two types of prompts: one for estimating CPT parameters, and another for personality trait intervention. Figure 4 displays the system prompt that we use to prompt the LLM to return the certainty equivalent to a prospect. Figure 5 shows the corresponding user prompt where that prompts the LLM with a specific prospect.

Experiment	Parameters	Value
CPT parameter estimates	System Prompt	120
	User Prompt	60
	Text-completion	200
	n. Models	6
	Runs	15
	Questions	56
	Total Tokens	2×10^6
Personified CPT parameter estimates	System Prompt	160
	User Prompt	60
	Text-completion	200
	n. Models	6
	Runs	15
	Questions	56
	Trait Levels	4
	Total Tokens	8.5×10^6
Personality tests	System Prompt	0
	User Prompt	90
	Text-completion	9
	n. Models	6
	Runs	15
	Total Tokens	2.5×10^6
Total tokens across all experiments		13×10^6

Table 4: Estimates of total token usage for text-completion APIs in each of our main experiments.

Package	Version	Reference
NumPy	2.0.1	Harris et al. (2020)
Pandas	2.2.2	pandas development team (2020); Wes McKinney (2010)
Python	3.10.6	–
SciPy	1.14.0	Virtanen et al. (2020)
Seaborn	0.13.2	Waskom (2021)

Table 5: Packages, Versions, and References used in our work.

Figure 6 shows the system prompt that we use to intervene on a personality trait of the LLM. In this example, we prompt the model to respond to subsequent tasks with low Conscientiousness (level 2).

Prospect user prompt

```
{{Outcome 1}} dollars with {{100 - p}}%
probability and {{Outcome 2}} dollars with
{{p}}% probability.
Let's think about this step by step
```

Figure 5: User prompt for a prospect presented to an LLM primed with the system prompt from Figure 4. Here, `Outcome 1` and `Outcome 2` represent the lowest and highest possible outcomes, respectively, and `p` is the percentage probability of `Outcome 2` occurring.

Personality intervention system prompt

```
For the following task, respond in a way
that matches this description:
"I'm very unsure, very messy, very
irresponsible, very lazy, very
undisciplined, very impractical, very
extravagant, very disorganized, very
negligent, very careless."
```

Figure 6: Example of the system prompt designed to intervene the Big Five personality trait of Conscientiousness at intensity level 2.

Prospect system prompt

```
You will be given a gamble.
What is the {{least positive / most
negative}} amount of money that you would
accept instead of taking the gamble?
Give your answer in the following format:
reason: reason for your choice answer:
amount
```

Figure 4: System prompt to estimate the certainty equivalent to a risky prospect. We define the certainty equivalent by asking for the `least positive` or `most negative` amount of money accepted depending on whether the expected value of the prospect is positive or negative respectively.

B.4 Big Five personality traits

Table 6 shows personality facets in the IPIP-NEO-300 inventory associated with each personality trait from the Big Five (Goldberg et al., 1999).

Table 7 shows the markers associated with each personality trait. These markers are utilized in the system prompt to intervene on the language model's personality traits.

Figure 7 illustrates the mapping from markers in Table 7 to personality trait levels. All markers for a given personality trait level are appended to each

```
levels = {
    1: f"extremely {low_marker}",
    2: f"very {low_marker}",
    3: f"{low_marker}",
    4: f"a bit {low_marker}",
    5: f"neither {low_marker} nor {
        high_marker}",
    6: f"a bit {high_marker}",
    7: f"{high_marker}",
    8: f"very {high_marker}",
    9: f"extremely {high_marker}",
}
```

Figure 7: Pythonic pseudocode demonstrating the mapping from personality prompt level to facet markers. All facet markers for each trait are given in Table 7.

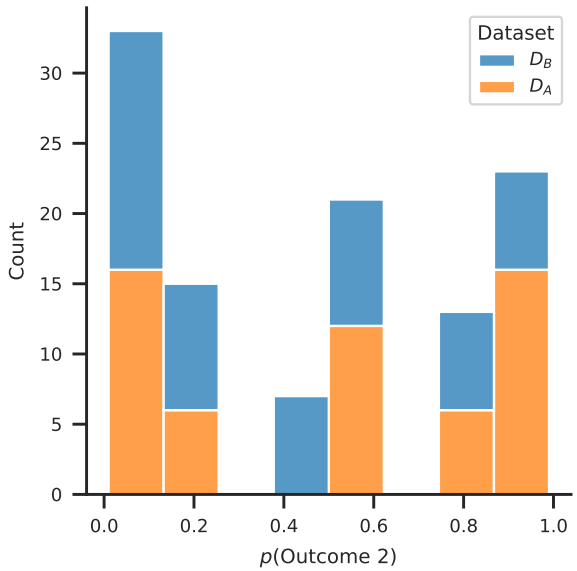
other for the intervention prompt. For example, *I'm very unsure, very messy, very irresponsible, very lazy, very undisciplined, very impractical, very extravagant, very disorganised, very negligent, very careless.* is the intervention for Conscientiousness at level 2.

Trait	Facets
OPE	Fantasy (O1), Aesthetics (O2), Feelings (O3), Actions (O4), Ideas (O5), Values (O6).
CON	Competence (C1), Order (C2), Dutifulness (C3), Achievement Striving (C4), Self-Discipline (C5), Deliberation (C6).
EXT	Warmth (E1), Gregariousness (E2), Assertiveness (E3), Activity (E4), Excitement-Seeking (E5), Positive-Emotions (E6).
AGR	Trust (A1), Straightforwardness (A2), Altruism (A3), Compliance (A4), Modesty (A5), Tender-Mindedness (A6).
NEU	Anxiety (N1), Angry Hostility (N2), Depression (N3), Self-Consciousness (N4), Impulsiveness (N5), Vulnerability (N6).

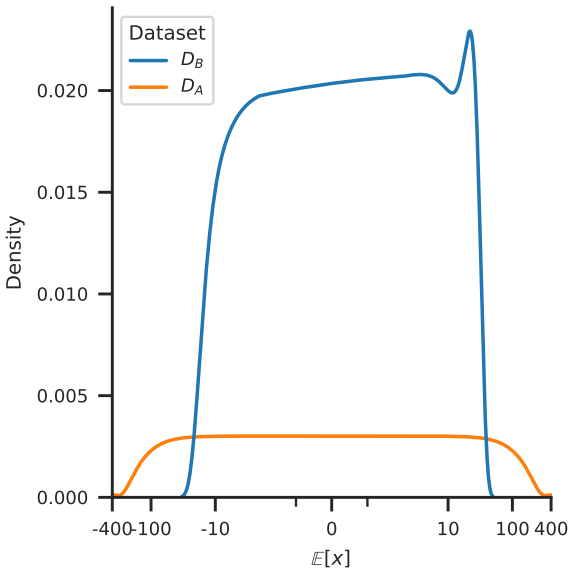
Table 6: Facets of the Big Five personality traits in the IPIP-NEO-300 personality inventory (Goldberg et al., 1999).

Trait	Low Marker	High Marker
Con.	unsure	self-efficacious
Con.	messy	orderly
Con.	irresponsible	responsible
Con.	lazy	hardworking
Con.	undisciplined	self-disciplined
Con.	impractical	practical
Con.	extravagant	thrifty
Con.	disorganised	organized
Con.	negligent	conscientious
Con.	careless	thorough
Ext.	unfriendly	friendly
Ext.	introverted	extraverted
Ext.	silent	talkative
Ext.	timid	bold
Ext.	unassertive	assertive
Ext.	inactive	active
Ext.	unenergetic	energetic
Ext.	unadventurous	adventurous and daring
Ext.	gloomy	cheerful
Agr.	distrustful	trustful
Agr.	immoral	moral
Agr.	dishonest	honest
Agr.	unkind	kind
Agr.	stingy	generous
Agr.	unaltruistic	altruistic
Agr.	uncooperative	cooperative
Agr.	self-important	humble
Agr.	unsympathetic	sympathetic
Agr.	selfish	unselfish
Agr.	disagreeable	agreeable
Neu.	relaxed	tense
Neu.	at ease	nervous
Neu.	easygoing	anxious
Neu.	calm	angry
Neu.	patient	irritable
Neu.	happy	depressed
Neu.	unselfconscious	self-conscious
Neu.	level-headed	impulsive
Neu.	contented	discontented
Neu.	emotionally stable	emotionally unstable
Ope.	unimaginative	imaginative
Ope.	uncreative	creative
Ope.	artistically unappreciative	artistically appreciative
Ope.	unaesthetic	aesthetic
Ope.	unreflective	reflective
Ope.	emotionally closed	emotionally aware
Ope.	uninquisitive	curious
Ope.	predictable	spontaneous
Ope.	unintelligent	intelligent
Ope.	unanalytical	analytical
Ope.	unsophisticated	sophisticated
Ope.	socially conservative	socially progressive

Table 7: Contrasting markers for low and high levels of the Big Five personality traits (Serapio-García et al., 2023). Low level markers run from 1–3 and high level markers run from 5–9.



(a) Distribution of probabilities associated with Outcome 2 for prospects in datasets D_A and D_B.



(b) Distribution of expected values for prospects in datasets D_A and D_B.

Figure 8: Distributions of (a) probabilities and (b) expected values for prospects in datasets D_A and D_B.

C Datasets

In this section, we provide a description of the data used in our work. We use two main prospect datasets:

- Dataset D_A: The 56 non-mixed prospects from Tversky and Kahneman (1992).
- Dataset D_B: The 56 mixed prospect randomly sampled from choices13k Peterson et al. (2021) used in the main body of our work (referred to as D in the main text).

Figure 8a and Figure 8b show the distributions of probabilities and expected values for prospects in both datasets, respectively. Table 8 provides a complete listing of all prospects in D_A and D_B.

Additionally, we utilize the IPIP-NEO-300 personality inventory data from Johnson (2020), which includes responses from 4808 individuals aged over 30 in the UK and Ireland. This data is publicly available and its distribution is permitted under the Open Science Framework's terms of use (Center for Open Science, 2025).

Outcome 1	**Outcome 2**	p	$\mathbb{E}[x]$	CE
0	50	0.10	5	9
0	50	0.50	25	21
0	50	0.90	45	37
0	−50	0.10	−5	−8
0	−50	0.50	−25	−21
0	−50	0.90	−45	−39
0	100	0.05	5	14
0	100	0.25	25	25
0	100	0.50	50	36
0	100	0.75	75	52
0	100	0.95	95	78
0	−100	0.05	−5	−8
0	−100	0.25	−25	−23.50
0	−100	0.50	−50	−42
0	−100	0.75	−75	−63
0	−100	0.95	−95	−84
0	200	0.01	2	10
0	200	0.10	20	20
0	200	0.50	100	76
0	200	0.90	180	131
0	200	0.99	198	188
0	−200	0.01	−2	−3
0	−200	0.10	−20	−23
0	−200	0.50	−100	−89
0	−200	0.90	−180	−155
0	−200	0.99	−198	−190
0	400	0.01	4	12
0	400	0.99	396	377
0	−400	0.01	−4	−14
0	−400	0.99	−396	−380
50	100	0.10	55	59
50	100	0.50	75	71
50	100	0.90	95	83
−50	−100	0.10	−55	−59
−50	−100	0.50	−75	−71
−50	−100	0.90	−95	−85
50	150	0.05	55	64
50	150	0.25	75	72.50
50	150	0.50	100	86
50	150	0.75	125	102
50	150	0.95	145	128
−50	−150	0.05	−55	−60
−50	−150	0.25	−75	−71
−50	−150	0.50	−100	−92
−50	−150	0.75	−125	−113
−50	−150	0.95	−145	−132
100	200	0.05	105	118
100	200	0.25	125	130
100	200	0.50	150	141
100	200	0.75	175	162
100	200	0.95	195	178
−100	−200	0.05	−105	−112
−100	−200	0.25	−125	−121
−100	−200	0.50	−150	−142
−100	−200	0.75	−175	−158
−100	−200	0.95	−195	−179

Dataset D_A.

Outcome 1	**Outcome 2**	p	$\mathbb{E}[x]$
29.00	37.00	0.05	29.40
16.00	47.00	0.50	31.50
−34.00	107.00	0.40	22.40
24.00	34.00	0.10	25.00
27.00	72.00	0.01	27.45
16.00	48.00	0.10	19.20
−14.00	37.00	0.10	−8.90
−19.00	0.00	0.95	−0.95
−16.00	16.00	0.80	9.60
2.00	90.00	0.01	2.88
−14.00	−3.00	0.05	−13.45
−28.00	38.00	0.60	11.60
3.00	26.00	0.25	8.75
−2.00	3.00	0.25	−0.75
−46.00	70.00	0.60	23.60
18.00	20.00	0.10	18.20
−23.00	24.00	0.99	23.53
−7.00	10.00	0.80	6.60
−5.00	−5.00	0.01	−5.00
−31.00	100.00	0.40	21.40
−21.00	36.00	0.25	−6.75
1.00	86.00	0.10	9.50
0.00	17.00	0.80	13.60
5.00	32.00	0.75	25.25
−12.00	58.00	0.60	30.00
−9.00	15.00	0.50	3.00
−7.00	35.00	0.50	14.00
−28.00	35.00	0.40	−2.80
−16.00	3.00	0.90	1.10
−2.00	90.00	0.25	21.00
−13.00	15.00	0.01	−12.72
−10.00	53.00	0.20	2.60
−10.00	29.00	0.99	28.61
−37.00	−8.00	0.90	−10.90
22.00	78.00	0.01	22.56
18.00	24.00	0.10	18.60
−23.00	82.00	0.40	19.00
−29.00	5.00	0.50	−12.00
−9.00	25.00	0.25	−0.50
−14.00	45.00	0.20	−2.20
0.00	68.00	0.40	27.20
9.00	11.00	0.10	9.20
11.00	14.00	0.90	13.70
−15.00	−4.00	0.75	−6.75
−14.00	53.00	0.25	2.75
−9.00	9.00	0.90	7.20
22.00	30.00	0.10	22.80
−16.00	11.00	0.20	−10.60
−24.00	28.00	0.40	−3.20
−3.00	21.00	0.40	6.60
−44.00	2.00	0.75	−9.50
−17.00	12.00	0.80	6.20
21.00	26.00	0.05	21.25
9.00	35.00	0.50	22.00
−9.00	70.00	0.50	30.50
−16.00	77.00	0.01	−15.07

Dataset D_B.

Table 8: Prospects in datasets D_A and D_B (D in the main text). Dataset D_A comprises all non-mixed prospects from Tversky and Kahneman (1992), and D_B comprises 56 mixed prospects randomly sampled from the choices13k dataset (Peterson et al., 2021). Outcome 1 and Outcome 2 are the dollar amounts for the risky prospect, $p = P(\text{Outcome 2})$ and CE is the prospect's certainty equivalent in D_B.

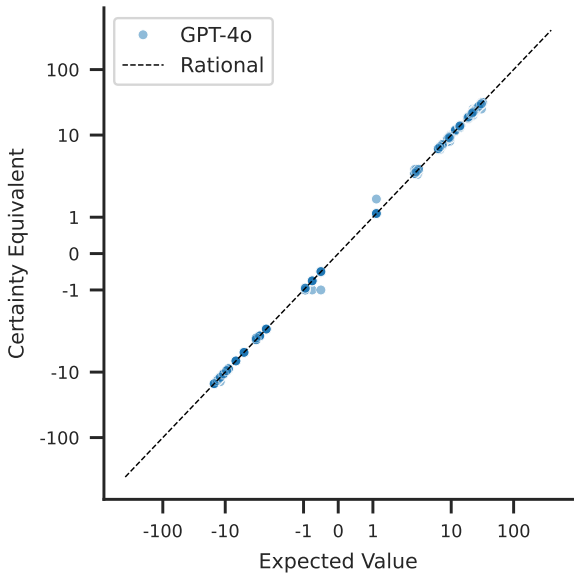


Figure 9: A comparison of the certainty equivalents for GPT-4o and the expected values of prospects in the dataset D_A using the system prompt from our work. The *Rational* data-series indicates the certainty equivalents of a risk-neutral rational agent.

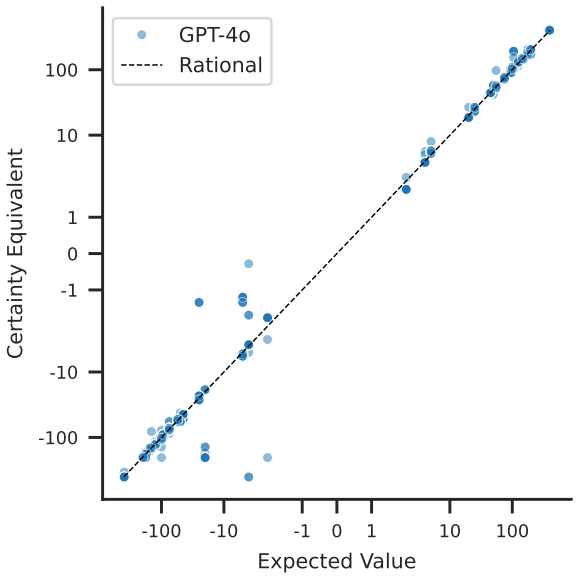


Figure 10: A comparison of the certainty equivalents for GPT-4o and the expected values of prospects in the dataset D_A using the system prompt from Ross et al. (2024). The *Rational* data-series indicates the certainty equivalents of a risk-neutral rational agent.

D Additional results

In this section we present results to support our work in the main body of the article.

D.1 System prompt ablations

In this section we show the difference in certainty equivalents estimated using our method and Ross et al. (2024) over prospects from D_A. Figure 9 and Figure 10 show a comparison of estimated certainty equivalents in our work and Ross et al. (2024) respectively over 15 experimental runs. Results for previous work show that certainty equivalents for prospects with potential losses are unstable. Where as the certainty equivalents for prospects with potential gains are stable. Our results are stable for prospects with potential gains and losses.

D.2 Certainty equivalents of LLMs for non-mixed gambles

We present a comprehensive analysis of decision-making behaviour in GPT-4o, directly comparing it with risk-neutral rational agents and 25 graduate students (Tversky and Kahneman, 1992) at the prospect level.

Figure 11 shows the certainty equivalents for GPT-4o, risk-neutral rational and human agents over a range of non-mixed prospects (D_A). We show the median certainty equivalents for each agent in Table 9. We use the methodology detailed in Section 3.1 Notably, GPT-4o exhibits certainty equivalents that are either identical or closely aligned with those of rational risk neural agents for nearly all prospects (i.e. certainty equivalents are equal to expected values of the prospects). This similarity suggests that GPT-4o's decision-making process closely adheres to the principles of expected utility theory, where the utility function is directly proportional to the outcome.

These findings extend the recent work by Ross et al. (2024), which demonstrated an increasing trend towards rationality in decision-making for GPT-4 and GPT-4-Turbo. Our results provide further evidence that more recent LLMs exhibit enhanced rational decision-making capabilities.

D.3 CPT parameters of LLMs on mixed-gambles

Table 10 presents the median values and 95% confidence intervals for the CPT model parameters measured in our study over mixed prospects D_B. For context, we include the median CPT parameter values for humans, as reported by Tversky and Kahneman (1992). It's important to note that confidence intervals are not available for the human data.

Outcome 1	Outcome 2	p	Median Certainty Equivalents (Rational/Human/GPT-4o)
0	50	0.1	5/9/5
0	50	0.5	25/21/25
0	50	0.9	45/37/45
0	−50	0.1	-5/-8/-5
0	−50	0.5	-25/-21/-25
0	−50	0.9	-45/-39/-45
0	100	0.05	5/14/5
0	100	0.25	25/25/25
0	100	0.5	50/36/50
0	100	0.75	75/52/75
0	100	0.95	95/78/95
0	−100	0.05	-5/-8/-5
0	−100	0.25	-25/-24/-25
0	−100	0.5	-50/-42/-50
0	−100	0.75	-75/-63/-75
0	−100	0.95	-95/-84/-95
0	200	0.01	2/10/2
0	200	0.1	20/20/20
0	200	0.5	100/76/100
0	200	0.9	180/131/180
0	200	0.99	198/188/198
0	−200	0.01	-2/-3/-2
0	−200	0.1	-20/-23/-20
0	−200	0.5	-100/-89/-100
0	−200	0.9	-180/-155/-180
0	−200	0.99	-198/-190/-198
0	400	0.01	4/12/4
0	400	0.99	396/377/396
0	−400	0.01	-4/-14/-4
0	−400	0.99	-396/-380/-396
50	100	0.1	55/59/55
50	100	0.5	75/71/75
50	100	0.9	95/83/95
−50	−100	0.1	-55/-59/-55
−50	−100	0.5	-75/-71/-75
−50	−100	0.9	-95/-85/-95
50	150	0.05	55/64/54
50	150	0.25	75/72/75
50	150	0.5	100/86/100
50	150	0.75	125/102/125
50	150	0.95	145/128/145
−50	−150	0.05	-55/-60/-55
−50	−150	0.25	-75/-71/-75
−50	−150	0.5	-100/-92/-100
−50	−150	0.75	-125/-113/-112
−50	−150	0.95	-145/-132/-145
100	200	0.05	105/118/105
100	200	0.25	125/130/125
100	200	0.5	150/141/150
100	200	0.75	175/162/175
100	200	0.95	195/178/195
−100	−200	0.05	-105/-112/-105
−100	−200	0.25	-125/-121/-125
−100	−200	0.5	-150/-142/-150
−100	−200	0.75	-175/-158/-175
−100	−200	0.95	-195/-179/-195

Table 9: Median certainty equivalents for Rational/Human/GPT-4o agents for non-mixed prospects from (Tversky and Kahneman, 1992). Outcome 1 and Outcome 2 are the dollar amounts for the risky prospect, and $p = P(\text{Outcome 2})$.

Model	θ	Median	CI lower	CI upper
Claude 3 Haiku	α	0.97	0.95	1.04
Claude 3 Haiku	β	0.97	0.94	1.11
Claude 3 Haiku	λ	1.19	0.82	1.53
Claude 3 Haiku	ϕ_+	0.93	0.88	1.00
Claude 3 Haiku	ϕ_-	0.86	0.67	1.04
Claude 3 Sonnet	α	1.00	0.99	1.02
Claude 3 Sonnet	β	0.99	0.97	1.01
Claude 3 Sonnet	λ	1.02	0.94	1.16
Claude 3 Sonnet	ϕ_+	0.98	0.95	1.03
Claude 3 Sonnet	ϕ_-	0.99	0.90	1.00
GPT-4o	α	0.99	0.99	1.00
GPT-4o	β	1.00	0.99	1.00
GPT-4o	λ	0.97	0.95	1.00
GPT-4o	ϕ_+	1.00	0.99	1.02
GPT-4o	ϕ_-	1.00	0.99	1.00
GPT-4o mini	α	0.94	0.90	0.99
GPT-4o mini	β	1.05	0.85	1.18
GPT-4o mini	λ	0.74	0.47	1.33
GPT-4o mini	ϕ_+	0.98	0.94	1.22
GPT-4o mini	ϕ_-	0.93	0.78	1.03
Gemini 1.5 Flash	α	1.00	1.00	1.00
Gemini 1.5 Flash	β	1.00	1.00	1.00
Gemini 1.5 Flash	λ	1.00	1.00	1.00
Gemini 1.5 Flash	ϕ_+	1.00	1.00	1.00
Gemini 1.5 Flash	ϕ_-	1.00	1.00	1.00
Gemini 1.5 Pro	α	1.00	1.00	1.00
Gemini 1.5 Pro	β	1.00	1.00	1.00
Gemini 1.5 Pro	λ	1.00	1.00	1.00
Gemini 1.5 Pro	ϕ_+	1.00	1.00	1.00
Gemini 1.5 Pro	ϕ_-	1.00	1.00	1.00
Human	α	0.88	0.00	0.00
Human	β	0.88	0.00	0.00
Human	λ	2.25	0.00	0.00
Human	ϕ_+	0.61	0.00	0.00
Human	ϕ_-	0.69	0.00	0.00

Table 10: Estimated CPT parameters for multiple LLMs. The table shows the median values and the upper and lower 95% confidence intervals (CI) for each parameter. Confidence intervals are calculated using bootstrapping with 10000 samples.

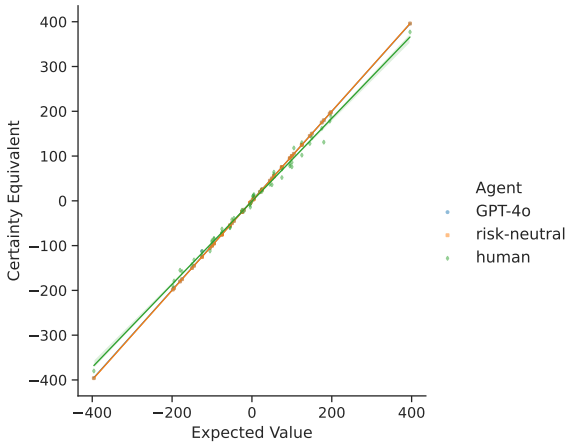


Figure 11: A comparison of certainty equivalents and expected values for non-mixed prospects from the dataset D_A across different agents. Linear regression lines for each agent are displayed, with shaded regions representing 95% confidence intervals. Note that that the results for GPT-4o and risk-neutral agents are near indistinguishable.

E Interventions on personality traits

This section demonstrates that trait interventions in GPT-4o and GPT-4o-mini produce clear variations in their corresponding trait scores measured using the BFI-44 personality inventory (John et al., 1999) while minimally affecting other Big Five personality traits. This confirms the causal relationship between interventions and personality trait scores as claimed in the main text.

We conducted an experiment with 50 runs for interventions on each personality trait over trait levels $\{1, 3, 7, 9\}$. For each intervention, we prompted the model to answer questions from the BFI-44 personality inventory. This approach validates the method proposed by Serapio-García et al. (2023) on an additional personality inventory beyond the IPIP-NEO-300.

We calculate the mean score for each level and measured trait across all interventions. Table 11 and Table 12 presents the minimum and maximum average scores for each measured trait and intervention for GPT-4o and GPT-4o-mini, respectively. The findings show that targeted personality interventions result in substantial variation in the intended trait while inducing minimal changes in non-targeted traits. For instance, in the case of GPT-4o, an intervention on Openness yields a full-range shift from a minimum score of 1 to a maximum of 5. In contrast, the non-targeted traits exhibit negligible differences between their minimum and maximum scores, suggesting high specificity of the intervention effect.

Table 13 and Table 14 present the statistical significance of one-tailed t-tests comparing personality scores between adjacent trait intervention levels. The results demonstrate that personality interventions produce monotonically increasing scores across all levels, with statistically significant differences between consecutive intervention strengths.

Target Trait	Measured Trait	Min. Score	Max score	Score range
OPE	**OPE**	1.0	5.0	**4.0**
OPE	CON	3.0	3.0	0.0
OPE	EXT	2.9	3.0	0.1
OPE	AGR	3.0	3.1	0.1
OPE	NEU	3.0	3.1	0.1
CON	OPE	2.9	3.0	0.1
CON	**CON**	1.3	5.0	**3.7**
CON	EXT	2.9	3.0	0.1
CON	AGR	3.0	3.1	0.1
CON	NEU	3.0	3.1	0.1
EXT	OPE	2.8	3.1	0.3
EXT	CON	3.0	3.0	0.0
EXT	**EXT**	1.0	5.0	**4.0**
EXT	AGR	2.8	3.3	0.5
EXT	NEU	2.7	3.1	0.4
AGR	OPE	3.0	3.0	0.0
AGR	CON	3.0	3.1	0.1
AGR	EXT	2.9	3.0	0.1
AGR	**AGR**	1.0	4.9	**3.9**
AGR	NEU	3.0	3.1	0.1
NEU	OPE	3.0	3.0	0.0
NEU	CON	3.0	3.0	0.0
NEU	EXT	2.9	3.0	0.1
NEU	AGR	2.7	3.8	1.1
NEU	**NEU**	1.0	5.0	**4.0**

Table 11: Range of GPT-4o's BFI-44 personality trait scores: minimum and maximum values for averaged personality scores at each level when systematically intervening on each target trait. Rows indicate the trait being manipulated (first column) and the trait being measured (second column).

Target Trait	Measured Trait	Min. Score	Max score	Score range
OPE	**OPE**	1.0	4.5	**3.5**
OPE	CON	3.1	3.2	0.1
OPE	EXT	2.2	3.2	1.0
OPE	AGR	3.1	3.5	0.4
OPE	NEU	2.8	3.0	0.2
CON	OPE	2.8	3.0	0.2
CON	**CON**	2.4	4.9	**2.5**
CON	EXT	2.9	3.0	0.1
CON	AGR	3.0	3.5	0.5
CON	NEU	2.7	3.2	0.5
EXT	OPE	2.6	3.2	0.6
EXT	CON	3.0	3.7	0.7
EXT	**EXT**	1.2	4.9	**3.7**
EXT	AGR	3.1	4.3	1.2
EXT	NEU	2.5	3.1	0.6
AGR	OPE	2.7	2.9	0.2
AGR	CON	2.9	3.5	0.6
AGR	EXT	2.8	3.0	0.2
AGR	**AGR**	1.8	4.8	**2.9**
AGR	NEU	2.9	3.3	0.4
NEU	OPE	2.9	2.9	0.0
NEU	CON	3.0	3.3	0.3
NEU	EXT	2.9	3.1	0.1
NEU	AGR	3.0	3.9	0.9
NEU	**NEU**	1.3	4.5	**3.2**

Table 12: Range of GPT-4o-mini's BFI-44 personality trait scores: minimum and maximum values for averaged personality scores at each level when systematically intervening on each target trait. Rows indicate the trait being manipulated (first column) and the trait being measured (second column).

Trait	Comparison	Significance
OPE	Level 3 score $>$ Level 1 score	***
OPE	Level 7 score $>$ Level 3 score	***
OPE	Level 9 score $>$ Level 7 score	***
CON	Level 3 score $>$ Level 1 score	***
CON	Level 7 score $>$ Level 3 score	***
CON	Level 9 score $>$ Level 7 score	***
EXT	Level 3 score $>$ Level 1 score	***
EXT	Level 7 score $>$ Level 3 score	***
EXT	Level 9 score $>$ Level 7 score	**
AGR	Level 3 score $>$ Level 1 score	***
AGR	Level 7 score $>$ Level 3 score	***
AGR	Level 9 score $>$ Level 7 score	***
NEU	Level 3 score $>$ Level 1 score	***
NEU	Level 7 score $>$ Level 3 score	***
NEU	Level 9 score $>$ Level 7 score	***

Table 13: One-tailed t-tests comparing BFI-44 personality trait scores across intervention levels $\{1, 3, 7, 9\}$ for GPT-4o. Results marked with */**/*** have statistical significance at $\alpha = 0.01/0.05/0.01$ level respectively.

Trait	Comparison	Significance
OPE	Level 3 score > Level 1 score	***
OPE	Level 7 score > Level 3 score	***
OPE	Level 9 score > Level 7 score	***
CON	Level 3 score > Level 1 score	***
CON	Level 7 score > Level 3 score	***
CON	Level 9 score > Level 7 score	***
EXT	Level 3 score > Level 1 score	**
EXT	Level 7 score > Level 3 score	***
EXT	Level 9 score > Level 7 score	***
AGR	Level 3 score > Level 1 score	*
AGR	Level 7 score > Level 3 score	***
AGR	Level 9 score > Level 7 score	***
NEU	Level 3 score > Level 1 score	**
NEU	Level 7 score > Level 3 score	***
NEU	Level 9 score > Level 7 score	***

Table 14: One-tailed t-tests comparing BFI-44 personality trait scores across intervention levels $\{1, 3, 7, 9\}$ for GPT-4o-mini. Results marked with */**/*** have statistical significance at $\alpha = 0.01/0.05/0.01$ level respectively.