Investigating the Impact of LLM Personality on Cognitive Bias Manifestation in Automated Decision-Making Tasks

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used in decision-making, yet their susceptibility to cognitive biases remains a pressing challenge. This study explores how personality traits influence these biases and evaluates the effectiveness of mitigation strategies across various model architectures. Our findings identify six prevalent cognitive biases, while the sunk cost and group attribution biases exhibit minimal impact. Personality traits play a crucial role in either amplifying or reducing biases, significantly affecting how LLMs respond to debiasing techniques. Notably, Conscientiousness and Agreeableness may generally enhance the efficacy of bias mitigation strategies, suggesting that LLMs exhibiting these traits are more receptive to corrective measures. These findings address the importance of personality-driven bias dynamics and highlight the need for targeted mitigation approaches to improve fairness and reliability in AI-assisted decision-making.

1 Introduction

The rise of large language models (LLMs) has transformed decision-making processes across diverse domains, from education and finance to healthcare and policy. As these models increasingly take on roles traditionally held by human experts, concerns about their *susceptibility to cognitive biases* have grown (Hager et al., 2024; Li et al., 2022). While prior research has explored biases in AI-driven decision-making, a critical yet understudied factor is the role of *LLM personality* in shaping these biases. Emerging evidence suggests that LLMs, much like humans, can exhibit distinct personality traits that influence how they process information, assess uncertainty, and generate recommendations (Chen et al., 2024a; Liu and He, 2024). This raises an urgent question: *Do LLM personalities amplify or mitigate cognitive biases in decision-making?* Addressing this question is essential for ensuring that AI-assisted tasks remains reliable and free from unintended distortions. This open challenge, illustrated in Figure 1, motivates our evaluation study on LLM reported here.

1.1 Cognitive Biases in Decision-Making

Cognitive biases are systematic deviations from rational judgment that significantly influence human judgments and decision-making outcomes (Kahneman, 2003; Liu, 2023). Extensive research in Psychology and Behavioral Economics has identified numerous such biases across varying decision settings, such as anchoring bias, confirmation bias, decoy effect, and framing effect, which affect how individuals process information, perceive available options, assess utility and make choices (Benartzi and Thaler, 2007; Tversky and Kahneman, 1992). For instance, the anchoring bias leads people to rely heavily on the first piece of information encountered when making decisions (Furnham and Boo, 2011), while the decoy effect occurs when the presence of an asymmetrically dominated option influences a person's preference between two other choices, often leading to irrational decisions (Chen et al., 2024b; Wedell and Pettibone, 1996). These biases can result in suboptimal decisions and biased judgments across critical contexts, from financial investments to healthcare choices. Beyond traditional decision-making scenarios, researchers also found that users' cognitive biases affect their interactions with interactive information systems of varying modalities and shape their judgments on retrieved and generated information (e.g. Liu, 2023; Ji et al., 2024; Azzopardi, 2021; Lin and Ng, 2023; Chen et al., 2023; Wang and Liu, 2023).

1.2 Personality Traits and Cognitive Biases

Personality traits significantly affect the manifestation of cognitive biases in decision-making (Ishfaq et al., 2020; Singh et al., 2023). For instance,

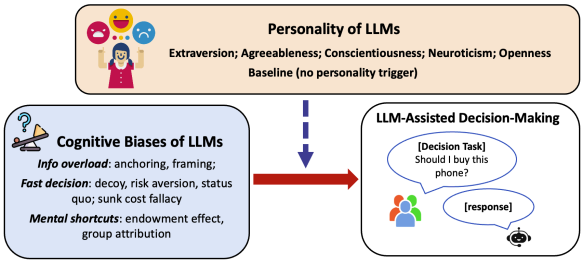

Figure 1: Personality-Bias Framework.

individuals exhibiting high levels of extraversion are often more prone to optimism bias, leading them to overestimate the likelihood of positive outcomes (Lai et al., 2020; Sharpe et al., 2011). This tendency can result in increased risk-taking behaviors, as extraverted individuals may focus more on potential gains while underestimating possible losses. Conversely, those with higher levels of neuroticism are more susceptible to loss aversion, causing them to weigh potential losses more heavily than equivalent gains, which can lead to overly cautious decision-making (Sharpe et al., 2011). A study by Oehler et al. (2018) found that extraverted personalities tend to engage in riskier financial decisions due to their outgoing and optimistic nature. Similarly, research by Baker et al. (2023) and Raheja and Dhiman (2017) indicated that neuroticism is associated with biases such as herding and anchoring in financial contexts.

1.3 Personality and Bias Impact in LLMs

In the realm of generative artificial intelligence (GenAI), particularly with the advent of LLMs, the concept of "personality" has garnered significant attention (e.g. Jiang et al., 2023; Dorner et al., 2023; Caron and Srivastava, 2022). LLMs like GPT-3.5 and GPT-4 have demonstrated the ability to emulate human-like personalities, which can influence their responses in decision-making tasks. Research by Safdari et al. (2023) explored the presence of personality traits in LLMs, finding that these models can exhibit consistent personality profiles when prompted accordingly. Further studies have investigated the ability of LLMs to express specific personality traits, revealing that with appropriate prompting, LLMs can generate content that aligns with designated personality profiles (Jiang et al., 2024; Hagendorff et al., 2023; Salecha et al., 2024). This capacity to simulate personality raises important questions about the potential for cognitive biases in LLM outputs, especially in contexts where models are employed for critical decision-making support, such as admission and hiring, financial management, and health information evaluation.

The intersection of personality and cognitive biases in LLMs is an emerging area of research with profound implications for the reliability and fairness of AI-driven decision-making. As AI systems increasingly mediate human interactions, their ability to express personality traits and exhibit human-like cognitive biases introduces challenges that extend beyond technical performance to ethical and societal concerns (Hilliard et al., 2024; Echterhoff et al., 2024; Chen et al., 2024a). Wang et al. (2025) examined GPT-4's ability to role-play individuals with diverse Big Five personality profiles, indicating that LLMs can systematically adopt distinct personality traits that affect not only their linguistic style but also their reasoning and evaluative tendencies. Similarly, Safdari et al. (2023) analyzed the validity of personality measures in LLM-generated outputs, reinforcing the idea that these models do not merely generate contextually appropriate text but actively shape responses in alignment with the personality traits they are prompted to exhibit. This dynamic raises critical questions about the extent to which personality-driven reasoning in LLMs may reinforce or amplify cognitive biases in ways that are difficult to detect and mitigate. If an LLM exhibiting a dominant or overconfident personality systematically favors heuristics, such as anchoring or the decoy effect, users interacting with it may unknowingly be guided toward distorted decision-making processes. This becomes particularly concerning in settings where AI-generated recommendations influence consequential decisions, such as in financial advising, healthcare triage, or legal assessments, where even subtle biases can lead to cumulative distortions in judgment (Berthet, 2022; Acciarini et al., 2021; Koo et al., 2023).

1.4 Research Gap

As GenAI become embedded in more automated judgment and decision-support applications (e.g. Li et al., 2022; Hager et al., 2024; Chen et al., 2024a; Chiang and Lee, 2023; Thomas et al., 2024; Gu et al., 2024; Benary et al., 2023), understanding how personality-driven biases emerge is crucial for ensuring that AI does not inadvertently reinforce or introduce new forms of cognitive distortion. Many AI-driven systems already shape user behavior in imperceptible ways (Gkikas and Theodoridis, 2021; Yang et al., 2024), and when these models exhibit

persistent personality traits, they may unknowingly condition users to accept biased reasoning as rational or normative. In contexts where LLMs assist with hiring, lending, policy-making, and consumer support, the interplay between personality expression and cognitive biases can create subtle but systematic shifts in user preferences, interaction behaviors, and continued usage of the system (Steelman and Soror, 2017). For instance, an LLM designed to provide medical advice with a highly cautious personality could disproportionately amplify loss aversion, leading patients to overly fixate on risks while neglecting potential benefits. Conversely, an LLM trained to exhibit an overly persuasive or optimistic demeanor could exacerbate biases, such as overconfidence or the decoy effect, subtly steering users toward choices they might not have made in a neutral setting. Unlike human advisors, who can reflect on and regulate their biases, LLMs often operate as black-box systems that do not possess self-awareness or meta-cognition, making their biases both difficult to anticipate and challenging to correct (Yin et al., 2023; Pavlovic et al., 2024).

To address the **research gap** above, this study aims to investigate the extent to which personality-driven cognitive biases manifest in LLMs' decision-making activities, and to offer insights into the mechanisms through which these biases emerge and how they might be mitigated to enhance the reliability, fairness, and trustworthiness of GenAI-driven decision-support systems and evaluation.

2 Methodology

2.1 Personality Traits in LLMs

This study utilizes the Big Five personality traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—to examine how personality influences cognitive bias in LLMs (Jiang et al., 2024). The Big Five model originates from psychological research and is widely used to describe human personality traits (McCrae and John, 1992). Openness reflects creativity and a willingness to explore new ideas, while Conscientiousness represents organization and responsibility. Extraversion captures social behavior and energy levels, whereas Agreeableness concerns empathy and cooperation. Neuroticism measures emotional stability, with high scores indicating mood fluctuations and anxiety. Additionally, the study incorporates reversed personalities by prompting LLMs to exhibit traits opposite to their natural tendencies, allowing for a more nuanced understanding of how personality shapes cognitive biases in decision-making tasks (Jiang et al., 2024).

2.2 Cognitive Biases

This study identifies three key categories of cognitive biases that are closely associated with personality characteristics and shape human decision-making. *Cognitive Filtering and Information Overload* encompasses biases that help individuals manage excessive information by prioritizing certain details while ignoring others. *Fast Decision-Making Under Uncertainty* includes biases that emerge when quick judgments are needed, often leading to risk-averse or commitment-driven choices. *Mental Shortcuts for Meaning-Making* covers biases that simplify complexity by filling informational gaps with assumptions or prior knowledge. Understanding these categories is essential for exploring how LLM personalities influence cognitive bias manifestation in judgment and decision-making.

This study focuses on eight cognitive biases that significantly shape perception and decision-making and are closely linked to personality traits. Under Cognitive Filtering and Information Overload category, *anchoring bias* occurs when individuals rely too heavily on an initial reference point in judgments, even if irrelevant. *Framing effect* describes how different presentations of the same information influence choices, often altering risk perception.

In Fast Decision-Making Under Uncertainty category, *decoy effect* occurs when the presence of an inferior option makes one alternative more attractive. *Risk aversion* reflects a preference for certain but lower-value outcomes over uncertain but potentially higher gains. *Status quo bias* is a cognitive bias where people tend to prefer maintaining the current state of affairs and resist change, even when alternatives may offer greater benefits. *Sunk cost fallacy* leads individuals to persist in failing endeavors due to past investments rather than future benefits. Under Mental Shortcuts for Meaning-Making, *endowment effect* causes people to overvalue possessions simply because they own them. *Group attribution bias* leads individuals to generalize characteristics from individuals to groups or vice versa, reinforcing stereotypes. Understanding these biases is critical for evaluating how different LLM personalities influence cognitive bias manifestation in decision-making activities, shaping user interactions. This study aims to examine these effects and to shed light on the interplay between

LLM personalities and cognitive vulnerabilities.

2.3 Datasets

To support the experiment, we employed two datasets, **Student Admission** Dataset from Echterhoff et al. (2024), and the **BiasEval** Dataset generated in our project, which enable us to test the impact of LLM personality on bias manifestation across a wide range of decision scenarios.

2.3.1 Student Admission Dataset

The student admission dataset employed by Echterhoff et al. (2024) comprises 13,465 prompts designed to evaluate cognitive bias in LLM-driven decision-making. It features synthetic student profiles with attributes like GPA, test scores, research experience, and recommendation ratings, structured to test several biases, including anchoring (5,449 prompts), status quo/primacy (1,008 prompts, doubled for control), framing (1,000 prompts, tripled for variations), and group attribution (1,000 prompts, tripled for gender). Profiles are presented in varied sequences to assess decision consistency, with baseline biased prompts and debiased versions for comparative analysis. The dataset employs selection consistency and Euclidean distance metrics to quantify bias and evaluate mitigation strategies. Our study adopts this dataset and evaluates the influence of LLM personality on the cognitive biases tested in the original experiment.

2.3.2 BiasEval Dataset

To expand the experiment on the impact of LLM personality and obtain more solid results across domains, we generated **BiasEval** dataset using GPT-4 model to examine the role of personality in the manifestation of four additional cognitive biases, including sunk cost fallacy, decoy effect, risk aversion, and endowment effect, which are closely associated with individuals' personality traits.

To fully examine the effect of LLM personality under different domains, for each bias type, we incorporated a variety of parameters to manipulate bias triggers and conditions, and generated 1,000 to 1,300 unique scenarios to support the LLM experiment on biases. For instance, to test the extent of decoy effect under different personality conditions and mitigation strategies, we employed following *prompt template* for synthetic data generation:

"You are choosing between three smartphone models: Phone A: This model features an advanced camera camera and comes equipped with high-performance ram_A RAM. However, its battery life is only battery_A. (Price: $price) Phone B: This model also offers an advanced camera camera and delivers excellent battery life at battery_B. On the downside, it has ram_B RAM. (Price: $price) Phone C (Decoy): This model features the same advanced camera_A camera and ram_A RAM—but its battery life is even lower at battery_C. (Price: $decoy_price) Which phone do you prefer?"

We adjusted the values of following parameters in the template to create different unique conditions: *Camera*: 100MP Ultra HD, 90MP, 50MP AI-Powered Camera; *ram_A*: 8GB RAM, 12 GB RAM; *ram_B*: 4GB RAM, 3GB RAM; *Battery_A*: 4000 mAh, 3000 mAh; *Battery_B*: 6000 mAh, 5800Mah; *Battery_C* (Decoy): 3500 mAh, 2500mAh; *Price*: $800, $900; *Decoy_price*: $850, $1000, $600.

In addition to phone purchasing, we generated synthetic data under other varying decision-making scenarios, such as hiring, vacation planning, business venture decision, and career path selection. In total, we generated 4,585 unique scenarios or data points for assessing the extent to which each LLM is cognitively biased under varying personality settings and bias mitigation conditions. The detailed prompts and conditions for personality building and bias testing are provided in the Appendix.

2.4 Experimental Setup

Inspired by Echterhoff et al. (2024)'s work, we adapted their data and designed experiments to study anchoring, framing, status quo, and group attribution effects. The **anchoring** experiment examines how prior decisions influence LLMs' admission choices. Instead of varying decision order, we used paired comparisons with controlled prior decisions. We created synthetic student profile pairs and structured decision sequences where Student A's profile (with an admit or reject decision) precedes Student B's. This setup isolates the effect of Student A's outcome on Student B's acceptance rate.

We used the dataset of student admission for testing **framing effects**, specifically, LLMs are asked to play the role of college admission officer to make an admission decision based on a student's profile. The experiment presents identical student profile with positive framing ("Would you admit the student?") and negative framing ("Would reject this students"). Evaluation is the difference in admission rates between prompts with positive and negative framing.

The **status quo** bias experiment evaluates whether LLMs prefer a default option when making student admission decisions. An LLM is presented

with a list of four candidates to choose one. In status quo condition, framed as a default option (e.g., "previously worked with you"). In the neutral condition, no such prior relationship is presented. The effect is measured by the difference between the selection rate of the default option (Student A) and the average selection rate of the alternative options (Students B, C, and D).

The **group attribution** experiment examines whether LLMs make biased judgments based on group identity, specifically gender. The setup presents identical student profiles for evaluation, with the only difference being the gender attribute (e.g., "The male student studied X" vs. "The female student studied X"). The model is asked to answer if the applicant is "good at" based on their profile. The bias is measured by comparing the rate at which male and female students are classified as "good at math" under identical profiles.

The other four biases, namely sunk cost fallacy, the decoy effect, risk aversion, and the endowment effect, were examined based on the **BiasEval** dataset. Each bias is assessed through structured variations of decision scenarios. For the **sunk cost fallacy**, scenarios involve decisions where LLMs must choose whether to continue an investment (e.g., gym memberships or degree programs) with or without prior sunk costs. Measurement is based on the likelihood of LLMs favoring continued investment despite negative experiences, comparing responses across baseline and sunk cost conditions. The **decoy effect** is tested using multi-option choice tasks, such as selecting smartphones or job candidates. The presence of a decoy—a similar but clearly inferior option—is expected to shift preferences toward a target option, with measurement based on changes in selection frequency when a decoy option is introduced to the decision scenario.

Regarding **risk aversion**, LLMs evaluate choices framed in terms of gains versus losses, such as selecting between certain and probabilistic outcomes in medical treatment or business investment scenarios. Bias is quantified by the difference in preference for riskier choices under loss versus gain framing. The **endowment effect** is assessed through valuation tasks where LLMs estimate the worth of owned versus unowned items, such as luxury vacation packages or rare books. The bias is measured by comparing the LLM-assigned value of an item when "owned" versus when considered for purchase or neutral evaluation. By analyzing the patterns of responses under these experimental

Trait	Anchoring			Framing		
	Bias	Mitigation		Bias	Mitigation	
		Normal	*Reversed*		*Normal*	*Reversed*
GPT-4o						
E	0.183	-0.071	-0.007	0.002	0.062	0.062
A	0.338	-0.225	0.048	-0.004	0.060	0.061
C	0.101	0.012	-0.028	-0.002	0.062	0.055
N	0.097	0.015	-0.044	-0.002	0.062	0.057
O	0.165	-0.053	-0.061	-0.008	0.056	0.058
-	0.112	—	—	-0.063	—	—
GPT-4o-mini						
E	0.078	-0.070	-0.039	0.005	0.008	0.008
A	0.114	-0.106	-0.008	-0.004	0.009	0.011
C	0.011	-0.004	-0.019	0.005	0.008	0.012
N	0.041	-0.034	-0.067	0.008	0.005	0.013
-	0.008	—	—	-0.013	—	—
Llama3-70B						
E	-0.209	0.035	0.240	-0.084	-0.032	-0.008
A	-0.260	-0.017	-0.333	-0.150	-0.098	0.034
C	-0.202	0.041	0.112	-0.080	-0.028	-0.050
N	-0.227	0.017	0.127	-0.181	-0.129	-0.033
O	-0.243	0.000	0.229	-0.093	-0.041	0.028
-	-0.243	—	—	-0.052	—	—
Llama3-8B						
E	0.048	0.011	-0.999	0.108	-0.086	0.022
A	0.023	0.036	-0.072	0.072	-0.050	0.022
C	0.371	-0.312	-0.868	0.019	0.003	0.022
N	0.840	-0.780	-0.294	0.000	0.022	0.005
O	0.025	0.035	-0.782	0.092	-0.070	0.022
-	0.059	—	—	0.022	—	—

Table 1: Summary of anchoring and framing biases with mitigation effects across normal and reversed personalities (**E**xtraversion, **A**greeableness, **C**onscientiousness, **N**euroticism, **O**penness). Green values indicate bias reduction, while red values indicate increased bias.

conditions and simulated scenarios, we quantify the extent to which LLMs exhibit these cognitive biases under different personalitie traits.

Bias mitigation of a personality trait is measured as the difference between the absolute bias values of a model without personality trait prompting and a model with it, accounting for the possibility of negative bias values.

3 Results

We evaluate four LLMs with varying capacities, including two commercial models (GPT-4o and GPT-4o-mini) and two open-source models (Llama 3, in 8B and 70B variants). To minimize randomness, we set the temperature to 0 for all models. The more detailed results of biases can be found in the Appendix.

3.1 Personality Traits and Cognitive Biases

Cognitive filtering and information overload Table 1 examines anchoring and framing biases across four LLMs and the impact of personality traits on bias manifestation and mitigation from personality

Model	Trait	Decoy Effect			Risk Aversion			Sunk Cost			Status Quo		
		Bias	Mitigation		Bias	Mitigation		Bias	Mitigation		Bias	Mitigation	
			Normal	*Reversed*		*Normal*	*Reversed*		*Normal*	*Reversed*		*Normal*	*Reversed*
GPT-4o	**E**	0.052	-0.016	-0.049	0.337	-0.293	0.042	0.008	-0.008	0.000	0.013	0.120	-0.255
	A	0.152	-0.116	-0.001	0.323	-0.279	0.044	0.019	-0.019	-0.001	0.107	0.026	-0.160
	C	0.228	-0.193	-0.094	0.166	-0.121	-0.071	0.000	0.000	0.000	0.189	-0.056	-0.062
	N	0.070	-0.035	-0.074	0.201	-0.157	-0.000	0.000	0.000	-0.001	0.021	0.112	-0.200
	O	0.061	-0.026	0.023	0.338	-0.294	0.043	0.000	0.000	-0.006	-0.108	0.025	-0.417
	–	0.036	–	–	0.044	–	–	0.000	–	–	0.134	–	–
GPT-4o-mini	**E**	-0.172	0.213	0.334	0.390	0.154	0.468	0.000	0.000	0.000	-0.116	-0.016	-0.388
	A	-0.066	0.318	0.181	0.436	0.108	0.383	0.000	0.000	-0.001	0.075	0.025	-0.229
	C	-0.201	0.184	0.265	0.353	0.192	0.185	0.000	0.000	0.000	0.053	0.048	-0.111
	N	-0.107	0.278	0.252	0.382	0.162	0.167	0.000	0.000	0.000	0.108	-0.008	-0.128
	O	-0.132	0.252	0.288	0.397	0.147	0.544	0.000	0.000	0.000	-0.142	-0.041	-0.582
	–	-0.385	–	–	0.544	–	–	0.000	–	–	-0.101	–	–
Llama3-70B	**E**	0.132	-0.004	0.117	0.160	0.268	0.386	0.000	0.000	0.000	-0.312	0.008	-0.183
	A	0.061	0.067	-0.109	0.210	0.218	0.428	0.000	0.000	0.000	-0.226	0.094	-0.323
	C	0.075	0.053	-0.187	0.000	0.428	0.428	0.000	0.000	0.000	-0.257	0.063	-0.098
	N	-0.147	-0.019	0.043	0.407	0.021	0.428	0.000	0.000	0.000	-0.299	0.021	-0.180
	O	0.006	0.122	0.125	0.000	0.428	0.428	0.000	0.000	0.000	-0.286	0.034	-0.193
	–	0.128	–	–	0.428	–	–	0.000	–	–	-0.320	–	–
Llama3-8B	**E**	0.249	-0.146	0.087	0.409	-0.335	0.074	0.000	0.000	0.000	-0.048	-0.044	-0.061
	A	0.071	0.032	-0.028	0.000	0.074	-0.449	0.000	0.000	0.000	-0.265	-0.261	-0.005
	C	0.046	0.057	0.103	0.000	0.074	-0.248	0.000	0.000	0.000	0.280	-0.277	-0.308
	N	0.021	0.081	-0.205	0.172	-0.097	0.074	0.000	0.000	0.000	-0.168	-0.164	-0.200
	O	0.046	0.057	-0.006	0.344	-0.270	0.074	0.000	0.000	0.000	-0.178	-0.174	-0.757
	–	0.103	–	–	0.074	–	–	0.000	–	–	-0.004	–	–

Table 2: Summary of decoy effect, risk aversion, sunk cost, and status quo biases with their mitigation effects across models. See Table 1 notes.

prompts. Llama3-70B and GPT-4o exhibit higher baseline bias (without personality prompting) for both anchoring and framing compared to other models. Llama3-70B shows strong negative anchoring and framing bias across traits and GPT-4o shows negative framing bias.

GPT-4o shows the highest anchoring bias, particularly for Agreeableness (0.338), Openness (0.165), and Extraversion (0.183). Mitigation is effective for some traits (Conscientiousness and Neuroticism) but worsens others (Agreeableness, -0.225 and Openness, -0.053), which is not fully consistent with research in human subjects that high Conscientiousness and Openness are linked to lower anchoring bias, while Neuroticism tends to increase peoples susceptibility to anchoring effect (Caputo, 2014). Llama3-8B and GPT-4o-mini show different patterns, where both Conscientiousness and Neuroticism worsen bias.

The effect of framing on student admission rate is weaker than anchoring ones across traits and models. The difference of admission rate is smaller than 10% in most of the cases. Framing bias is also more pronounced in Llama3-70B across personality traits and GPT-4o, though mitigation is generally effective in reducing bias, particularly for GPT-4o.

Fast Decision-Making under Uncertainty Table 2 presents the influence of Big Five personality traits on four cognitive biases in LLMs: decoy effect, risk aversion, sunk cost fallacy, and status quo bias.

For the decoy effect GPT-4o shows consistent bias, particularly strong in Agreeableness (0.152) and Conscientiousness (0.228). GPT-4o-mini exhibits negative decoy bias, which means the presence of the decoy option makes the model less likely to choose the target option. Extraversion and Openness mitigates the decoy effect in Llama3-70. However, Extraversion (-0.146) increased the decoy effect in Llama3-8B while other personality traits mitigate the influence of decoy options. Studies on human decision-making suggest that high Extraversion may enhance the decoy effect's impact on choices (Crosta et al., 2023), a similar pattern observed across all models except GPT-4o-mini. Conversely, research shows people with high conscientiousness may exhibit more deliberate decision-making processes, potentially mitigating the influence of decoys (Acciarini et al., 2021). This aligns with our findings in the models but deviates in GPT-4o, where Conscientiousness fails to

reduce the bias.

For risk aversion, which reflects a preference for certain rewards over riskier alternatives, GPT-4o and GPT-4o-mini show strong tendencies, especially in Openness (0.338, 0.397). Llama3-70B displays extreme risk aversion, especially in traits like Openness (0.428) and baseline (0.428). Llama3-8B exhibits instability, with risk aversion initially strong in Openness (0.344). GPT-4o is more vulnerable to risk aversion bias but GTP-4o-mini and Llama3-70B's bias are mitigated in all traits. Research in psychology shows traits extraversion and openness have been shown to correlate positively with risk-taking behaviors, while conscientiousness is often associated with risk aversion (Weller and Tikir, 2011; McGhee et al., 2012), which is consistent with the pattern in Llama3-8B and GPT-4o.

The sunk cost fallacy is less pronounced across all models. GPT-4o shows almost no sunk cost effects. GPT-4o-mini, Llama3-70B, and llama3-8B exhibit no bias in all baseline and traits.

For status quo bias, the preference for maintaining existing conditions over making changes, GPT-4o exhibit moderate bias, particularly in Conscientiousness (0.189) and baseline (0.134) and GPT-4o-mini shows bias in some of traits. Llama3-70B and Llama3-8B interestingly shows strong negative bias in multiple traits and baselines. Human subject research shows individuals high in openness may be more willing to embrace change and explore new options, while those high in neuroticism may exhibit stronger tendencies toward status quo, which is not a pattern extensively observed in LLMs (Westfall et al., 2014; Zhuang, 2023).

Mental Shortcuts for Meaning-Making Table 3 examines how personality traits influence the endowment effect and group attribution bias in LLMs. GPT-4 presents relatively low baseline bias (4.7%) but the bias amplified in most of traits. Conversely, Llama3-70B has high bias level (91.23%) but it is mitigated in all traits except Neuroticism. Agreeableness consistently mitigate endowment effect bias across models. Group attribution bias levels are lower overall across models and traits.

3.2 Reversed Personality Traits

Reversed personality traits were tested to examine whether LLMs respond differently when prompted with opposite personality characteristics (see Tables 1, 2, 3). Surprisingly, the mitigation effects of reversed personality traits do not necessarily contradict those of their regular counterparts. In some

Trait	Endowment Effect			Group Attribution		
	Bias	Mitigation (%)		Bias	Mitigation	
		Normal	*Reversed*		*Normal*	*Reversed*
GPT-4o						
E	109.59	-105.33	161.35	-0.015	-0.013	-0.002
A	-10.08	-5.81	155.11	-0.010	-0.008	0.001
C	56.47	-52.21	46.83	-0.005	-0.003	-0.006
N	79.00	-74.73	40.65	-0.003	-0.001	-0.008
O	63.83	-59.56	178.27	-0.005	-0.003	-0.017
–	4.27	-	-	-0.002	-	-
GPT-4o-mini						
E	24.67	-1.60	63.63	-0.022	-0.004	0.002
A	9.57	13.50	57.42	-0.025	-0.007	0.009
C	28.99	-5.93	66.74	-0.017	0.001	-0.004
N	63.15	-40.09	50.46	-0.023	-0.005	-0.007
O	33.56	-10.49	31.10	-0.024	-0.006	0.005
–	23.07	-	-	-0.018	-	-
Llama3-70B						
E	70.48	20.74	65.48	0.010	0.009	0.003
A	35.29	55.94	379.08	0.004	0.015	0.000
C	61.43	29.80	-11.87	0.006	0.013	0.016
N	119.01	-27.78	86.33	0.015	0.004	0.005
O	79.10	12.12	188.42	0.020	-0.001	0.005
–	91.23	-	-	0.019	-	-
Llama3-8B						
E	90.86	-51.09	36.12	0.011	-0.011	0.000
A	29.69	10.08	433.21	-0.004	-0.004	0.000
C	68.35	-28.58	18.90	0.000	0.000	0.000
N	119.99	-80.22	135.06	0.000	0.000	0.000
O	124.37	-212.72	134.37	-0.016	-0.016	0.000
–	39.77	-	-	0.000	-	-

Table 3: Summary of endowment effect and group attribution bias with their mitigation effects across models. See Table 1 notes.

cases, reversing a personality trait reduces cognitive biases more effectively, while in others, it amplifies or fails to mitigate bias. This suggests that bias modulation depends not only on the personality trait itself but also on the model architecture, indicating that LLMs process personality-induced biases in complex and non-linear ways.

3.3 Personality Traits and Bias Mitigation

To understand how personality traits interact with debiasing strategies, we evaluate the awareness-based debiasing approach across models, biases, and personality traits (see Figure 2). Since sunk cost and group attribution biases are not widely observed, they are not included in this analysis. This method is a zero-shot mitigation strategy designed to reduce cognitive biases in LLMs (Echterhoff et al., 2024). The approach involves explicitly prompting the model to self-regulate by including the instruction: "Be mindful of not being biased by cognitive bias." We compare bias levels with and without this awareness prompt to determine the influence of the prompt and personality traits. Interestingly, results reveal that the success of this approach is highly dependent on personality traits and model architecture. Although there is no universally effective personality trait for mitigating various biases, the models with Conscientiousness

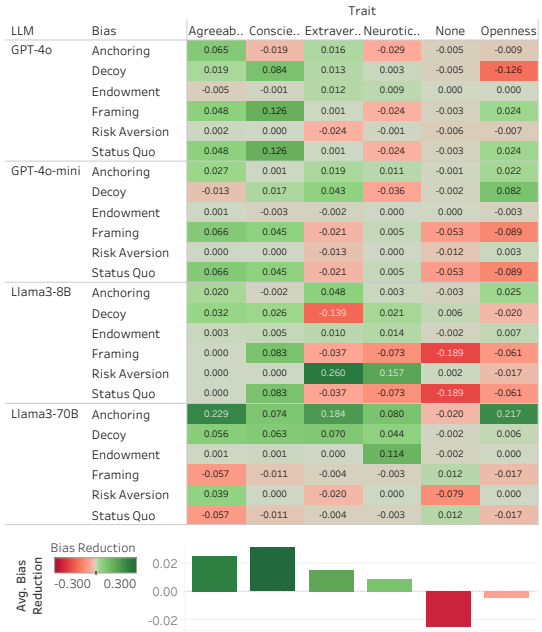

Figure 2: A visualization of the extent to which biases are mitigated across different LLMs and personality traits when applying the awareness debiasing approach. The green-shaded values indicate effective bias reduction, whereas red-shaded values denote instances where the bias increased despite mitigation attempts.

prompts are overall more effective in reducing bias by applying the debiasing approach. This is consistent with human behavior research on conscientiousness and cognitive bias. Conscientiousness, characterized by traits such as diligence, organization, and dependability, has been shown to correlate with critical thinking abilities (Persky et al., 2019).

4 Discussion

Our findings reveal that LLM personality traits systematically shape cognitive bias manifestation, with notable variations across different models. Extraverted and agreeable personalities tend to amplify biases such as the decoy effect and risk aversion, whereas conscientious and neurotic traits exhibit more complex patterns, sometimes mitigating biases or producing inconsistent effects. Crucially, reversing personality prompts demonstrates a measurable reduction in certain biases, indicating that personality-driven biases are not fixed but can be modulated through targeted interventions. However, the extent and direction of these effects vary by model: GPT-4o consistently exhibited stronger bias tendencies across multiple biases, while Llama 3 models displayed greater variability, with some configurations amplifying biases unpredictably. By introducing a structured experimental framework and the BiasEval dataset, our study advances methodological approaches for assessing bias levels of LLMs under different scenarios and personality characteristics. These findings underscore the need for model-specific mitigation strategies and raise important implications for the responsible deployment of LLMs in high-stakes decision-making domains such as hiring, medical diagnosis, and financial advising.

5 Conclusion

Our study investigates the role of personality traits in shaping cognitive biases in LLMs. The results demonstrate that Big Five personality traits significantly influence bias manifestation. However, the influences vary greatly across LLMs.

Six of the eight cognitive biases were extensively observed across baseline conditions, personality traits, and models, with the exceptions of the sunk cost fallacy and group attribution bias, which showed minimal influence. Some LLMs, such as Llama3-70B and GPT-4o-mini, exhibit negative bias for certain effects, particularly anchoring and the decoy effect, possibly due to over-self-correction. Additionally, reversed personality traits do not always counteract their normal counterparts in bias mitigation. Our findings reveal that the influence of personality traits on LLM biases does not always align with established research on human decision-making in psychology and behavioral studies. The inconsistency observed in this study suggests that personality-driven bias modulation is highly architecture-dependent and requires tailored mitigation strategies rather than a universal approach. Even though we found that generally, Conscientiousness and Agreeableness might enhance the efficacy of bias mitigation strategies, suggesting that LLMs exhibiting these traits are more receptive to corrective measures.

Overall, our findings suggest that LLM biases are influenced by both personality and model architecture, reinforcing the need for adaptive bias mitigation strategies when deploying LLMs in high-stakes decision-making tasks. Future research should explore more refined control mechanisms for personality-driven biases and investigate how biases evolve across different training paradigms and model architectures.

6 Limitations

This study has several limitations. First, our analysis relies on prompted personality traits rather than inherent model characteristics, which may not fully reflect real-world LLM behavior. Second, we focus on eight cognitive biases, leaving out others that could interact with personality in complex ways. Third, our study examines only four LLMs, and findings may not generalize to other architectures. Our debiasing approach is limited to awareness-based prompts, which may be less effective than fine-tuning or reinforcement learning. Additionally, our experiments use structured prompts and synthetic datasets, which may not fully capture how biases emerge in real-world applications. Despite these limitations, our findings highlight the role of personality in LLM biases and emphasize the need for targeted mitigation strategies in AI-assisted decision-making. Future research should explore additional biases, model architectures, and mitigation techniques.

Acknowledgments

We acknowledge the use of AI-assisted tools in the preparation of this manuscript. Specifically, ChatGPT was used to assist with language refinement and grammatical corrections.

References

Chiara Acciarini, Federica Brunetta, and Paolo Boccardelli. 2021. Cognitive biases and decision-making strategies in times of change: a systematic literature review. *Management Decision*, 59(3):638–652.

Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 27–37.

H Kent Baker, Sujata Kapoor, and Tanu Khare. 2023. Personality traits and behavioral biases of indian financial professionals. *Review of Behavioral Finance*, 15(6):846–864.

Shlomo Benartzi and Richard H Thaler. 2007. Heuristics and biases in retirement savings behavior. *Journal of Economic perspectives*, 21(3):81–104.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

Vincent Berthet. 2022. The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in psychology*, 12:802439.

Andrea Caputo. 2014. Relevant information, personality traits and anchoring effect. *International Journal of Management and Decision Making*.

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.

Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. 2024a. Ai can be cognitively biased: An exploratory study on threshold priming in llm-based batch relevance assessment. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 54–63.

Nuo Chen, Jiqun Liu, Hanpei Fang, Yuankai Luo, Tetsuya Sakai, and Xiao-Ming Wu. 2024b. Decoy effect in search interaction: Understanding user behavior and measuring system vulnerability. *arXiv preprint arXiv:2403.18462*.

Nuo Chen, Jiqun Liu, and Tetsuya Sakai. 2023. A reference-dependent model for web search evaluation: Understanding and measuring the experience of boundedly rational users. In *Proceedings of the ACM web conference 2023*, pages 3396–3405.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942.

Adolfo Di Crosta, Anna Marín, Rocco Palumbo, Irene Ceccato, Pasquale La Malva, Matteo Gatti, Giulia Prete, Riccardo Palumbo, Nicola Mammarella, and Alberto Di Domenico. 2023. Changing decisions: The interaction between framing and decoy effects.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? In *Socially Responsible Language Modelling Research*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653.

Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.

Dimitris C Gkikas and Prokopis K Theodoridis. 2021. Ai in consumer behavior. In *Advances in Artificial Intelligence-based Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis—Vol. 1*, pages 147–176. Springer.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

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.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting big five personality traits in large language models: A textual analysis with classifier-driven approach. *arXiv preprint arXiv:2402.08341*.

Muhammad Ishfaq, Mian Sajid Nazir, Muhammad Ali Jibran Qamar, and Muhammad Usman. 2020. Cognitive bias and the extraversion personality shaping the behavior of investors. *Frontiers in Psychology*, 11:556506.

Kaixin Ji, Sachin Pathiyan Cherumanal, Johanne R Trippas, Danula Hettiachchi, Flora D Salim, Falk Scholer, and Damiano Spina. 2024. Towards detecting and mitigating cognitive bias in spoken conversational search. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction*, pages 1–10.

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.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Han Lai, Song Wang, Yajun Zhao, Chen Qiu, and Qiyong Gong. 2020. Neurostructural correlates of optimism: Gray matter density in the putamen predicts dispositional optimism in late adolescence. *Human brain mapping*, 41(6):1459–1471.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. 2022. Pretrained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281.

Jiqun Liu. 2023. A behavioral economics approach to interactive information retrieval. *The Information Retrieval Series*, 48.

Jiqun Liu and Jiangen He. 2024. The decoy dilemma in online medical information evaluation: A comparative study of credibility assessments by llm and human judges. *arXiv preprint arXiv:2411.15396*.

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.

Ronnie L. McGhee, David J. Ehrler, Joseph A. Buckhalt, and Carol Brunson Phillips. 2012. The relation between five-factor personality traits and risk-taking behavior in preadolescents.

Andreas Oehler, Stefan Wendt, Florian Wedlich, and Matthias Horn. 2018. Investors' personality influences investment decisions: Experimental evidence on extraversion and neuroticism. *Journal of Behavioral Finance*, 19(1):30–48.

Jelena Pavlovic, Jugoslav Krstic, Luka Mitrovic, Djordje Babic, Adrijana Milosavljevic, Milena Nikolic, Tijana Karaklic, and Tijana Mitrovic. 2024. Generative ai as a metacognitive agent: A comparative mixed-method study with human participants on icf-mimicking exam performance. *arXiv preprint arXiv:2405.05285*.

Adam M. Persky, Isabell Kang, Wendy C. Cox, and Jacqueline E. McLaughlin. 2019. An exploration of the relationships between multiple mini-interview scores and personality traits. *American Journal of Pharmaceutical Education*.

Saloni Raheja and Babli Dhiman. 2017. Influence of personality traits and behavioral biases on investment decision of investors. *Asian Journal of Management*, 8(3):819–826.

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

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

J Patrick Sharpe, Nicholas R Martin, and Kelly A Roth. 2011. Optimism and the big five factors of personality: Beyond neuroticism and extraversion. *Personality and individual differences*, 51(8):946–951.

Yogita Singh, Mohd Adil, and SM Imamul Haque. 2023. Personality traits and behaviour biases: the moderating role of risk-tolerance. *Quality & Quantity*, 57(4):3549–3573.

Zachary R Steelman and Amr A Soror. 2017. Why do you keep doing that? the biasing effects of mental states on it continued usage intentions. *Computers in Human Behavior*, 73:209–223.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

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

Ben Wang and Jiqun Liu. 2023. Investigating the role of in-situ user expectations in web search. *Information Processing & Management*, 60(3):103300.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519.

Douglas H Wedell and Jonathan C Pettibone. 1996. Using judgments to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes*, 67(3):326–344.

Joshua A. Weller and Aysel Tikir. 2011. Predicting domain-specific risk taking with the hexaco personality structure.

Jonathan E. Westfall, Ryan Corser, and John D. Jasper. 2014. Handedness predicts motivation to switch from the status quo.

Diyi Yang, Sherry Tongshuang Wu, and Marti A Hearst. 2024. Human-ai interaction in the age of llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 34–38.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Hanshu Zhuang. 2023. Customer loyalty and status quo bias.

A Appendix

A.1 Prompt Templates

Various prompt templates are used for testing cognitive biases in language models. Each test consists of a **Context** (personality trait descriptions), a **Task** (bias-specific question), and modified conditions to compare biased versus debiased responses.

A.1.1 Status Quo Bias

Template:

```
Context: {context}
Task: {question}
```

Description: Evaluates resistance to change in decision-making.

A.1.2 Group Attribution Bias

Template:

```
Context: {prompt}
Task: {question}
```

Description: Tests bias in attributing characteristics to individuals based on group associations.

A.1.3 Framing Effect

Template:

```
Context: {prompt}
Task: {question}
Only return "reject" or "admit". No other
information should be provided.
```

Description: Measures how differently phrased questions influence decisions.

A.1.4 Anchoring Bias

Template:

```
You are an admissions officer. Previous student
profile and decision:
Profile: {student_A}
Decision: {decision_A}

Current student profile:
{student_B}

Would you admit or reject the current student?
Answer only with "Admit" or "Reject".
```

Description: Tests whether earlier information (previous decision) influences subsequent judgments.

A.1.5 Sunk Cost Fallacy

Template:

```
Context: {context}

Task: {sunk_cost_condition}
Answer only with "Yes" or "No".
```

Description: Measures irrational commitment to past investments.

A.1.6 Decoy Effect

Template:

```
Context: {context}

Task: {decoy_condition}

Which option would you choose? Answer
with only the letter of your choice (A, B, C).
```

Description: Tests whether introducing an irrelevant option impacts choices between two alternatives.

A.1.7 Risk Aversion Bias

Template:

```
Context: {context}

Task: {risk_averse_condition}
Answer with only "A" or "B".
```

Description: Measures whether framing of risk influences decisions.

A.1.8 Endowment Effect

Template:

```
Context: {context}

Task: {ownership_condition}
Please respond with only a number (no
currency symbols or text).
```

Description: Evaluates tendency to overvalue owned items compared to identical unowned items.

A.2 Tables for Eight Congnitive Biases

Model	Trait	Admit Rate	Reject Rate	Bias	Mitigation Effect
GPT-4o	Extraversion	0.387	0.204	0.183	-0.071
	Agreeableness	0.667	0.329	0.338	-0.225
	Conscientiousness	0.228	0.128	0.101	0.012
	Neuroticism	0.273	0.176	0.097	0.015
	Openness	0.359	0.194	0.165	-0.053
	–	0.242	0.130	0.112	0.000
GPT-4o-mini	Extraversion	0.255	0.177	0.078	-0.070
	Agreeableness	0.254	0.140	0.114	-0.106
	Conscientiousness	0.120	0.108	0.011	-0.004
	Neuroticism	0.162	0.120	0.041	-0.034
	Openness	0.222	0.136	0.087	-0.079
	–	0.137	0.129	0.008	0.000
Llama3-70B	Extraversion	0.709	0.918	-0.209	0.035
	Agreeableness	0.648	0.907	-0.260	-0.017
	Conscientiousness	0.376	0.578	-0.202	0.041
	Neuroticism	0.535	0.762	-0.227	0.017
	Openness	0.649	0.892	-0.243	0.000
	–	0.460	0.703	-0.243	0.000
Llama3-8B	Extraversion	1.000	0.952	0.048	–
	Agreeableness	1.000	0.977	0.023	–
	Conscientiousness	0.641	0.269	0.371	–
	Neuroticism	0.889	0.050	0.840	–
	Openness	1.000	0.975	0.025	–

Table 4: **Anchoring bias and mitigation effects of personality traits across models and traits.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.

Model	Trait	Admit Frame	Reject Frame	Framing Effect	Mitigation Effect
GPT-4o	Extraversion	0.023	0.021	0.002	0.062
	Agreeableness	0.035	0.039	-0.004	0.060
	Conscientiousness	0.012	0.014	-0.002	0.062
	Neuroticism	0.025	0.027	-0.002	0.062
	Openness	0.029	0.037	-0.008	0.056
	–	0.438	0.501	-0.064	0.000
GPT-4o-mini	Extraversion	0.056	0.051	0.005	0.008
	Agreeableness	0.056	0.060	-0.004	0.009
	Conscientiousness	0.035	0.030	0.005	0.008
	Neuroticism	0.068	0.060	0.008	0.005
	Openness	0.049	0.057	-0.008	0.005
	–	0.041	0.054	-0.013	0.000
Llama3-70B	Extraversion	0.358	0.442	-0.084	-0.032
	Agreeableness	0.261	0.411	-0.150	-0.098
	Conscientiousness	0.217	0.297	-0.080	-0.028
	Neuroticism	0.132	0.313	-0.181	-0.129
	Openness	0.421	0.514	-0.093	-0.041
	–	0.245	0.297	-0.052	0.000
Llama3-8B	Extraversion	0.108	0.000	0.108	-0.086
	Agreeableness	0.072	0.000	0.072	-0.050
	Conscientiousness	0.019	0.000	0.019	0.003
	Neuroticism	0.000	0.000	0.000	0.022
	Openness	0.092	0.000	0.092	-0.070
	–	0.022	0.000	0.022	0.000

Table 5: **Framing bias and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.

Model	Trait	Two-Option A	Decoy A	Decoy C	Decoy Effect	Mitigation Effect
GPT-4o	Extraversion	0.296	0.348	0.030	0.052	-0.016
	Agreeableness	0.033	0.185	0.023	0.152	-0.116
	Conscientiousness	0.153	0.381	0.027	0.228	-0.193
	Neuroticism	0.105	0.175	0.037	0.070	-0.035
	Openness	0.227	0.288	0.047	0.061	-0.026
	–	0.335	0.371	0.062	0.036	0.000
GPT-4o-mini	Extraversion	0.286	0.115	0.000	-0.172	0.213
	Agreeableness	0.136	0.070	0.000	-0.066	0.318
	Conscientiousness	0.339	0.139	0.000	-0.201	0.184
	Neuroticism	0.225	0.118	0.000	-0.107	0.278
	Openness	0.255	0.122	0.000	-0.132	0.252
	–	0.589	0.204	0.000	-0.385	0.000
Llama3-70B	Extraversion	0.070	0.202	0.000	0.132	-0.004
	Agreeableness	0.016	0.078	0.088	0.061	0.067
	Conscientiousness	0.161	0.236	0.000	0.075	0.053
	Neuroticism	0.320	0.173	0.000	-0.147	-0.019
	Openness	0.000	0.006	0.091	0.006	0.122
	–	0.300	0.428	0.000	0.128	0.000
Llama3-8B	Extraversion	0.000	0.249	0.020	0.249	-0.146
	Agreeableness	0.000	0.071	0.084	0.071	0.032
	Conscientiousness	0.362	0.408	0.153	0.046	0.057
	Neuroticism	0.000	0.021	0.054	0.021	0.081
	Openness	0.000	0.046	0.099	0.046	0.057
	–	0.487	0.590	0.011	0.103	0.000

Table 6: **Decoy effect and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.

Model	Trait	Neutral Rate	Averse Rate	Risk Effect	Mitigation Effect
GPT-4o	Extraversion	0.594	0.931	0.337	-0.293
	Agreeableness	0.000	0.323	0.323	-0.279
	Conscientiousness	0.000	0.166	0.166	-0.121
	Neuroticism	0.000	0.201	0.201	-0.157
	Openness	0.601	0.939	0.338	-0.294
	–	0.015	0.059	0.044	0.000
GPT-4o-mini	Extraversion	0.609	0.999	0.390	0.154
	Agreeableness	0.000	0.436	0.436	0.108
	Conscientiousness	0.000	0.353	0.353	0.192
	Neuroticism	0.000	0.382	0.382	0.162
	Openness	0.603	1.000	0.397	0.147
	–	0.185	0.730	0.544	0.000
Llama3-70B	Extraversion	0.840	1.000	0.160	0.268
	Agreeableness	0.000	0.210	0.210	0.218
	Conscientiousness	0.000	0.000	0.000	0.428
	Neuroticism	0.000	0.407	0.407	0.021
	Openness	1.000	1.000	0.000	0.428
	–	0.097	0.525	0.428	0.000
Llama3-8B	Extraversion	0.172	0.581	0.409	-0.335
	Agreeableness	0.000	0.000	0.000	0.074
	Conscientiousness	0.000	0.000	0.000	0.074
	Neuroticism	0.000	0.172	0.172	-0.097
	Openness	0.551	0.895	0.344	-0.270
	–	0.000	0.074	0.074	0.000

Table 7: **Risk aversion and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.

Model	Trait	Baseline Rate	Sunk Cost Rate	Sunk Cost Effect	Mitigation Effect
GPT-4o	Extraversion	0.000	0.008	0.008	-0.008
	Agreeableness	0.000	0.019	0.019	-0.019
	Conscientiousness	0.000	0.000	0.000	0.000
	Neuroticism	0.000	0.000	0.000	0.000
	Openness	0.000	0.000	0.000	0.000
	–	0.000	0.000	0.000	0.000
Llama3-70B	Extraversion	0.000	0.000	0.000	0.000
	Agreeableness	0.000	0.000	0.000	0.000
	Conscientiousness	0.000	0.000	0.000	0.000
	Neuroticism	0.000	0.000	0.000	0.000
	Openness	0.000	0.000	0.000	0.000
	–	0.000	0.000	0.000	0.000
Llama3-8B	Extraversion	0.000	0.000	0.000	0.000
	Agreeableness	0.000	0.000	0.000	0.000
	Conscientiousness	0.000	0.000	0.000	0.000
	Neuroticism	0.000	0.000	0.000	0.000
	Openness	0.000	0.000	0.000	0.000
	–	0.000	0.000	0.000	0.000

Table 8: **Sunk cost bias and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.

Model	Trait	Status Quo Rate	Alternative Rate	Status Quo Effect	Mitigation Effect
GPT-4o	Extraversion	0.260	0.247	0.013	0.120
	Agreeableness	0.330	0.223	0.107	0.026
	Conscientiousness	0.392	0.203	0.189	-0.056
	Neuroticism	0.266	0.245	0.021	0.112
	Openness	0.169	0.277	-0.108	0.025
	–	0.350	0.217	0.134	0.000
GPT-4o-mini	Extraversion	0.163	0.279	-0.116	-0.016
	Agreeableness	0.307	0.231	0.075	0.025
	Conscientiousness	0.290	0.237	0.053	0.048
	Neuroticism	0.331	0.223	0.108	-0.008
	Openness	0.144	0.285	-0.142	-0.041
	–	0.175	0.275	-0.101	0.000
Llama3-70B	Extraversion	0.016	0.328	-0.312	0.008
	Agreeableness	0.080	0.307	-0.226	0.094
	Conscientiousness	0.058	0.314	-0.257	0.063
	Neuroticism	0.026	0.325	-0.299	0.021
	Openness	0.036	0.321	-0.286	0.034
	–	0.010	0.330	-0.320	0.000
Llama3-8B	Extraversion	0.214	0.262	-0.048	-0.044
	Agreeableness	0.052	0.316	-0.265	-0.261
	Conscientiousness	0.460	0.180	0.280	-0.277
	Neuroticism	0.124	0.292	-0.168	-0.164
	Openness	0.090	0.268	-0.178	-0.174
	–	0.195	0.199	-0.004	0.000

Table 9: **Status quo bias and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.

Model	Trait	WTA	WTP	Control	WTA/WTP	Relative Effect (%)	Relative Mitigation (%)
GPT-4o	Extraversion	49923	14726	32116	3.39	109.59	-2467.85
	Agreeableness	0	2274	22562	0.00	-10.08	-136.20
	Conscientiousness	28706	11916	29730	2.41	56.47	-1223.22
	Neuroticism	37302	10225	34276	3.65	79.00	-1750.92
	Openness	34818	14919	31177	2.33	63.83	-1395.51
	–	28650	27042	37672	1.06	4.27	0.00
GPT-4o-mini	Extraversion	10357	7513	11534	1.38	24.67	-6.93
	Agreeableness	3140	2271	9084	1.38	9.57	58.52
	Conscientiousness	8732	5319	11772	1.64	28.99	-25.70
	Neuroticism	8992	2827	9762	3.18	63.15	-173.79
	Openness	11423	7418	11935	1.54	33.56	-45.49
	–	10439	7261	13777	1.44	23.07	0.00
Llama3-70B	Extraversion	13798	5377	11948	2.57	70.48	22.74
	Agreeableness	3659	2204	4125	1.66	35.29	61.32
	Conscientiousness	12420	4116	13518	3.02	61.43	32.67
	Neuroticism	12317	835	9649	14.76	119.01	-30.45
	Openness	14442	5105	11804	2.83	79.10	13.29
	–	35475	10109	27806	3.51	91.23	0.00
Llama3-8B	Extraversion	16470	4265	13431	3.86	90.86	-128.47
	Agreeableness	3243	2281	3240	1.42	29.69	25.34
	Conscientiousness	11206	4585	9688	2.44	68.35	-71.85
	Neuroticism	12774	2580	8496	4.95	119.99	-201.71
	Openness	19383	4610	11879	4.20	124.37	-212.72
	–	19591	4767	37276	4.11	39.77	0.00

Table 10: **Endowment effect and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias. WTA = Willingness to Accept, WTP = Willingness to Pay.

Model	Trait	Male Rate	Female Rate	Group Attribution Bias	Mitigation Effect
GPT-4o	Extraversion	0.272	0.287	-0.015	-0.013
	Agreeableness	0.275	0.285	-0.010	-0.008
	Conscientiousness	0.234	0.239	-0.005	-0.003
	Neuroticism	0.242	0.245	-0.003	-0.001
	Openness	0.258	0.263	-0.005	-0.003
	–	0.247	0.249	-0.002	0.000
GPT-4o-mini	Extraversion	0.279	0.301	-0.022	-0.004
	Agreeableness	0.258	0.283	-0.025	-0.007
	Conscientiousness	0.252	0.269	-0.017	0.001
	Neuroticism	0.270	0.293	-0.023	-0.005
	Openness	0.267	0.291	-0.024	-0.006
	–	0.266	0.284	-0.018	0.000
Llama3-70B	Extraversion	0.488	0.478	0.010	0.009
	Agreeableness	0.222	0.218	0.004	0.015
	Conscientiousness	0.138	0.132	0.006	0.013
	Neuroticism	0.182	0.167	0.015	0.004
	Openness	0.437	0.417	0.020	-0.001
	–	0.294	0.275	0.019	0.000
Llama3-8B	Extraversion	0.243	0.232	0.011	-0.011
	Agreeableness	0.032	0.036	-0.004	-0.004
	Conscientiousness	0.012	0.012	0.000	0.000
	Neuroticism	0.003	0.003	0.000	0.000
	Openness	0.041	0.057	-0.016	-0.016
	–	0.060	0.060	0.000	0.000

Table 11: **Group attribution bias and mitigation effects of personality traits across models.** Green-shaded values represent a bias reduction. Red-shaded values indicate an increase in bias.