# BehaviorBench: A Psychologically Grounded Benchmark for Evaluating Personality in Large Language Models through Realistic Behaviors

**Taowen Pu**[1,2], **Hexi Wang**[3,2], **Zeyang Liu**[4,2], **Dongsheng Guo**[2*] and **Chuan Zhao**[1,2]
[1]Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, Jinan, China
[2]Quan Cheng Laboratory, Jinan, China
[3]Tsinghua University, Beijing, China
[4]Shandong University, Jinan, China
sr-guodsh@qcl.edu.cn

## Abstract

Current approaches to evaluating personality in large language models (LLMs) typically prompt them to self-report on psychological questionnaires such as the Big Five Inventory. However, these methods assess introspective labels rather than observable behavior, despite the fact that LLMs are deployed to act in realistic contexts, not to reflect on their own traits. To bridge this gap, we introduce BehaviorBench, a new benchmark to evaluate LLM personality through concrete behaviors in everyday scenarios. Grounded in established personality psychology, BehaviorBench links each Big Five trait to validated behavioral manifestations and embeds them in contextually plausible situations that naturally elicit trait-relevant actions. We evaluate a range of models and personality shaping strategies using BehaviorBench and find a substantial mismatch between self-reported personalities and actual behaviors. By grounding evaluation in observable behavior rather than introspection, our work reveals critical gaps in current LLM personality modeling and control mechanisms. Our code and data are available at https://github.com/butra1n/BehaviorBench

## 1 Introduction

Large language models (LLMs) have exhibited human-like personality traits, even in the absence of explicit design [Miotto *et al.*, 2022; Karra *et al.*, 2022]. This emerging anthropomorphism has spurred significant interest in *personality shaping* [Caron and Srivastava, 2023], which is the deliberate alignment of an LLM's output with specific personality profiles, to enhance performance in applications such as role-playing agents [Zhang *et al.*, 2025] and social simulation [Tu *et al.*, 2023]. However, evaluating these shaping methods requires reliable frameworks to ensure their validity and effectiveness [Han *et al.*, 2025].

Current evaluation paradigms predominantly adopt human psychological instruments, such as the Big Five Inventory (BFI) or Myers-Briggs Type Indicator (MBTI) questionnaires, by prompting LLMs to self-report their agreement with trait-descriptive statements [Huang *et al.*, 2023b]. Although convenient, this approach is fundamentally flawed: in real-world applications, LLMs are expected to perform tasks and make behavioral decisions in context. However, these scales are designed to measure human personality through introspection of traits, rather than through concrete behaviors. Humans possess stable internalized personalities that ensure the coherence between self-reported traits and actual behaviors. In contrast, when LLMs self-report traits via scales, the explicit relationship between scale items and trait-based prompts changes the task to instruction-following. However, when making behavioral decisions, the implicit trait-behavior relationships require LLMs to perform the analysis. Therefore, evaluating personality through behavioral decisions can better assess the ability of LLMs to analyze such relationships and simulate the behaviors of specific personalities. Such inherent flaws create a mismatch between LLMs' self-reported traits and their real behavioral outputs, undermining the validity of the existing assessment paradigm [Han *et al.*, 2025].

To address this, research has shifted the focus from labels to behavioral outputs, designing tasks to elicit and analyze LLM actions as proxies for personality [Lee *et al.*, 2025;

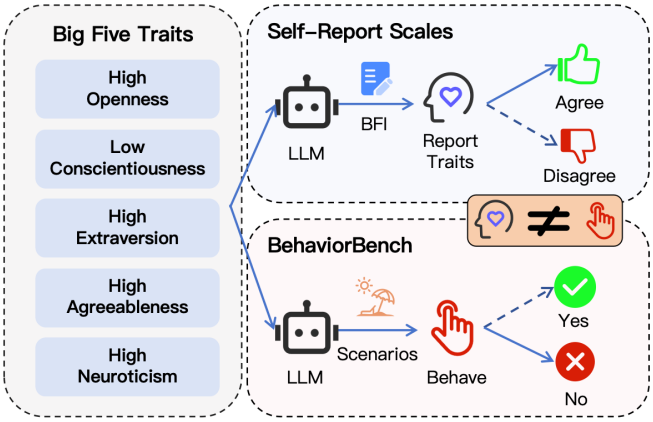


Figure 1: Comparison between trait-based self-report scales and BehaviorBench. BehaviorBench grounds each Big Five trait in psychologically validated, behaviorally rich scenarios that simulate real-world contexts and elicit trait-relevant behavioral decisions.

*Corresponding author.

Hartley *et al.*, 2025; Han *et al.*, 2025]. Although promising, existing behavior-centric approaches suffer from two limitations: weak theoretical grounding in trait-behavior mappings and low ecological validity of artificial scenarios, which undermine their reliability for real-world applications.

To overcome these limitations, we propose a psychologically grounded framework for LLM personality evaluation that assesses personality through behavior in realistic contexts. Our core insight is to leverage well-established, replicated findings from personality psychology that link each Big Five trait to specific, measurable behavioral indicators. We operationalize these links by constructing diverse everyday scenarios that naturally elicit the target behaviors, thereby ensuring both theoretical validity and ecological relevance. Based on this, we propose **BehaviorBench**, a benchmark dataset comprising 510 carefully designed behavior-scenario pairs across Big Five personality traits (*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*) [John *et al.*, 1991]. Each item presents a concrete real-world situation and defines a clear binary behavioral response aligned with psychological theory. To evaluate personality expression, we introduce two complementary metrics, i.e., the behavioral endorsement rate, which measures the frequency of trait-consistent behaviors, and the personality influence score, which quantifies the deviation from a neutral baseline in the theoretically expected direction. Based on these two indicators, BehaviorBench evaluates theoretically predicted directional changes in validated behaviors induced by personality shaping. Fig. 1 illustrates the distinction between BehaviorBench and trait-based self-report scales. Using BehaviorBench, we conduct systematic experiments across multiple LLMs and personality shaping strategies. The results show mainly that (1) quantified trait prompts yield significantly more consistent behaviors than descriptive ones; (2) LLMs exhibit a strong default bias toward high Agreeableness, making low-Agreeableness traits hard to elicit without contextual anchoring; (3) Neuroticism remains largely unresponsive to prompting, revealing a fundamental gap in modeling mood-related behaviors; (4) there is a discrepancy between self-reported personality scores and behavior, undermining the validity of trait-based self-report evaluation; and (5) LLM reasoning can enhance alignment between target traits and observed behaviors.

In summary, our main contributions are as follows:

- We introduce BehaviorBench, a benchmark for evaluating personality expression in LLMs via behavioral decisions. Unlike other methods, we build a psychologically grounded mapping that links each Big Five trait to validated behavioral indicators within realistic scenarios.
- We construct a comprehensive dataset that ensures high ecological validity for personality assessment. This behavior-driven approach reflects personality traits more faithfully than previous methods, being closer to real-world application scenarios.
- We conduct extensive experiments across various LLMs and prompting strategies to analyze behavioral expression. Our analysis reveals the gap between models' self-reported traits and behaviors, offering new insights into the consistency of personality shaping in LLMs.

## 2 Related Works

### 2.1 Personality Shaping of LLMs

For personality shaping, current approaches mainly use psychologically grounded prompts, such as the Big Five [McCrae and Costa, 1987] and the Myers-Briggs Type Indicator (MBTI) [Briggs, 1976], to induce target personalities in LLMs [Tu *et al.*, 2023; Guo *et al.*, 2025; Besta *et al.*, 2025]. Prompt strategies include two types: (1) specifying personality types or traits (e.g., "act as an INTJ" or "be highly extraverted") [Besta *et al.*, 2025; Deng *et al.*, 2025] and (2) defining traits with fine-grained descriptions. For instance, Jiang *et al.* [2023] directly generate natural-language statements that characterize high and low levels of each Big Five dimension, while Liu *et al.* [2024] quantify traits on discrete scales and incorporates the scores explicitly as contextual tokens. Some works also internalize personality via fine-tuning, as in BIG5-CHAT [Li *et al.*, 2025]. However, existing methods fail to link abstract traits to context-sensitive behaviors, causing LLMs to deviate from assigned personalities in practice [Han *et al.*, 2025].

### 2.2 Personality Assessments for LLMs

Personality psychology offers a framework for modeling individual differences via well-established trait models such as the Big Five [John *et al.*, 1991]. It is commonly used in LLM personality assessment by prompting models to respond to Big Five Inventory (BFI) items on a Likert scale of 1-5 [John *et al.*, 1991; Huang *et al.*, 2023b]. Recent studies show a mismatch between LLMs' self-reported traits and behavioral outputs, challenging the validity of this paradigm [Zou *et al.*, 2024; Han *et al.*, 2025; Huang *et al.*, 2023a]. Alternative approaches elicit open-ended responses and infer traits via human or LLM raters [Wang *et al.*, 2024; Jiang *et al.*, 2023], better capturing behavioral expression but suffering from ad hoc trait-behavior mappings and high annotation costs. Other works use controlled tasks to probe personality through behavior. For example, Hartley *et al.* [2025] link Big Five traits to risk-taking in gambling scenarios via prospect theory [Kahneman and Tversky, 2013]; Han *et al.* [2025] design tasks on risk, bias, honesty, and sycophancy to expose gaps between self-expression and action; and Lee *et al.* [2025] pair trait scales with ATOMIC10$\times$ [West *et al.*, 2022] scenarios to elicit behaviorally grounded choices. Despite progress, existing approaches face three limitations: (1) use of artificial scenarios lacking ecological validity, (2) limited task diversity and behavioral specificity, and (3) weak theoretical grounding in personality psychology, especially regarding the justification of trait-to-behavior mappings.

## 3 BehaviorBench

BehaviorBench is a behavior-based framework grounded in empirical trait–behavior mappings from personality psychology. Fig. 2 outlines BehaviorBench: (1) linking Big Five traits to behavior categories based on psychological foundations, (2) constructing a "trait-behavior-scenario" dataset, and (3) evaluating personality shaping.

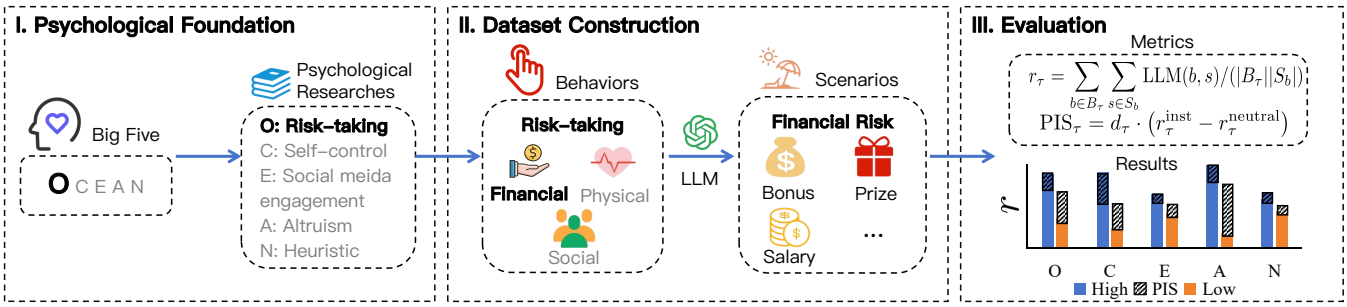


Figure 2: Overview of the BehaviorBench. Based on psychological research, we select behaviors related to each trait of the Big Five and generate diverse real-world scenarios for specific behaviors. In these scenarios, by testing the behavioral performance of LLMs and calculating the behavioral endorsement rate ($r$) and personality influence score (PIS), we can analyze the effect of LLMs' personality shaping.

### 3.1 Psychological Foundations

Personality psychology examines systematic individual differences in thoughts, affect, and behavior [Roberts and Yoon, 2022]. Among trait models, the Big Five have received the strongest empirical support due to its replicability, cross-cultural validity, and robust links to observable behavior [John *et al.*, 1991; Cobb-Clark and Schurer, 2012]. While alternative frameworks such as MBTI [Briggs, 1976] and Dark Triad [Paulhus and Williams, 2002] exist, the Big Five remains the dominant paradigm in behavioral research.

Critically, extensive evidence shows that Big Five traits are not only temporally stable but also reliably manifest in specific behavioral patterns across contexts [Al Otaibi, 2012; McElroy and Dowd, 2007; DeYoung *et al.*, 2002]. A core principle of our evaluation design is that personality can be inferred from observable actions rather than self-report, leveraging the well-established correspondence between traits and behavior. For each dimension, we identify one core behavioral construct with strong empirical backing:

**Openness → Risk-taking.** Individuals high in openness show curiosity and willingness to engage in novel experiences, leading to increased risk-taking across financial, social, and physical domains [Tok, 2011; Kipman *et al.*, 2021].

**Conscientiousness → Self-control.** Conscientious individuals are disciplined, exhibiting superior impulse regulation and persistence in goal-directed tasks, even under frustration or distraction [Jensen-Campbell *et al.*, 2007; Mao *et al.*, 2018].

**Extraversion → Social Media Engagement.** Extraverts are outgoing and sociable, actively participating in digital social environments through content creation, interaction, and feature adoption [Meng and Leung, 2021; Breil *et al.*, 2019].

**Agreeableness → Altruism toward Acquaintances.** Agreeable individuals are helpful and unselfish, displaying prosocial behavior, particularly toward friends or familiar others, highlighting the relational boundaries of altruism [Ashton *et al.*, 1998; Oda *et al.*, 2014].

**Neuroticism → Heuristic Reliance.** Neurotic individuals tend to be easily nervous and handle stress badly, so they are more prone to heuristic decision-making, including representativeness heuristic (judging by typical feature matching), availability heuristic (relying on ease of recall), and anchoring adjustment heuristic (using an initial value as a reference point), to alleviate anxiety under stress [Neijenhuis and Kupiainen, 2023; Tversky and Kahneman, 1974; Belhekar, 2017; Hilbig, 2008].

These mappings form the theoretical backbone of our evaluation framework and adhere to core psychometric principles: by isolating individual traits and adopting a theoretically validated behavioral proxy, we can precisely measure directional behavioral shifts induced by personality shaping, while eliminating confounding effects arising from overlapping trait motivations and contextual biases.

### 3.2 Dataset Construction

We operationalize the above trait–behavior links into a structured dataset $D = \{(\tau, b, s_b)\}$, where $\tau$ is a Big Five trait, $b$ is an empirically validated behavior associated with $\tau$, and $s_b$ is a realistic scenario in which $b$ may manifest (see Table 1 for examples). Formally, the dataset is defined as

$$D = \{d \mid d = (\tau, b, s_b),\ \tau \in T,\ b \in B_\tau,\ s_b \in S_b\},$$

where $T$ denotes the set of Big Five personality traits, $B_\tau$ is the set of behaviors associated with trait $\tau$, and $S_b$ is the set of scenarios corresponding to behavior $b$.

**Step 1: Selecting behaviors.** We draw from established psychological scales and experiments while applying two filters: (1) excluding legally or ethically sensitive actions (e.g., drug use) to avoid model refusals; (2) removing or merging redundant items describing the same underlying behavior (e.g., adding "music" or "subtitles" to videos). Specifically:

- *Risk-taking* (Openness): 10 behaviors (3 financial, 3 social, 4 physical) from DOSPERT [Blais and Weber, 2006].
- *Self-control* (Conscientiousness): 5 high self-controlled behaviors and 5 low self-controlled contrast cases from the Self-Control Scale [Tangney *et al.*, 2018].
- *Social media engagement* (Extraversion): 11 behaviors (5 interaction, 2 tool-use, 4 content-creation) based on Meng *et al.* [2021].
- *Altruism* (Agreeableness): 10 prosocial behaviors toward acquaintances from the Self-Report Altruism Scale [Rushton *et al.*, 1981], following Oda *et al.* [2014].
- *Heuristics* (Neuroticism): Two canonical behaviors (representativeness heuristic and availability heuristic) based on Tversky and Kahnema [1974]. Anchoring

| |
|---|
| **Openness** [Investing 10% of your annual income in a moderate growth mutual fund.]: Your tax return was processed this week, resulting in a $4,200 refund deposit. This amount equals approximately 10% of your previous year's income. |
| **Conscientiousness** [Keeping on schedule.]: You set aside 10:00 AM to 12:00 PM for deep work on a project. A friend messages asking to meet for lunch at 10:30 AM. |
| **Extraversion** [Live streaming.]: You are practicing a musical instrument alone in a soundproof room at a community center. |
| **Agreeableness** [Giving money to a friend who needed it (or asked me for it).]: During a team lunch meeting, a coworker realizes they left their wallet at home and need to cover their $15 meal share. |
| **Neuroticism** [Bicycle or motorcycle]: Statistics show that riding a bicycle in the city has fewer fatal accidents per mile than riding a motorcycle. However, you recently saw multiple news reports about bicycle accidents involving celebrities. Which mode of transportation would you choose for your daily commute. |

Table 1: Dataset instances from BehaviorBench, formatted as *Trait [Behavior (or Choices for Neuroticism)]: Scenario*

heuristic is excluded due to interference from LLM knowledge.

The list of behaviors is provided in the supplementary file.

**Step 2: Generating scenarios.** For each behavior, we use Qwen3-235B to generate realistic, diverse and neutral scenarios adhering to three principles: (1) depicting common daily scenarios; (2) ensuring diversity across scenarios for the same behavior; (3) including only objective facts with no suggestive language. We apply a one-shot prompting strategy and generate (all prompts can be found in the supplementary file):

- 10 scenarios per behavior for risk-taking, self-control, social media, and altruism;
- 50 scenarios per heuristic type (representativeness, availability), generated with formal definitions, paradigmatic examples, and detailed explanations thereof.

All generated scenarios are manually reviewed independently by three expert annotators with backgrounds in psychology and NLP, with disagreements resolved through discussion. Any scenarios that fail to satisfy the three principles are discarded and regenerated by the model until they meet the criteria, meaning that no formal inter-annotator metrics are adopted.

Notably, each instance $d \in D$ is associated with a binary judgment task: given the scenario $s_b$, a model's response is recorded as either "yes" or "no" regarding whether it endorses the behavior $b$. Across multiple samples, a consistent tendency to select "yes" indicates alignment with the behavior and thus reflects a higher level of the associated trait (e.g., high Openness for risk-taking), whereas a predominant preference for "no" suggests rejection of the behavior and corresponds to a lower level of that trait (e.g., low Openness). For Neuroticism, however, we adopt a forced-choice format in which the model selects between two mutually exclusive responses; one option is pre-designated as the behaviorally indicative label aligned with the heuristic. Choosing this label is treated as equivalent to a "yes" response. The forced-choice paradigm operationalizes heuristic reliance as the preference between a rational and a heuristic-consistent option under uncertainty. This preference structure cannot be reduced to a binary endorsement, such as "bicycle vs motorcycle". This design enables the quantification of behavioral tendencies through aggregate response statistics, providing a measurable basis for evaluating the effectiveness of personality shaping in LLMs.

In contrast to Han *et al.* [2025], who focus on detecting personality hallucination through behavioral inconsistency, BehaviorBench implements a comprehensive trait–behavior assessment protocol. Whereas TRAIT [Lee *et al.*, 2025] aims to develop a general-purpose personality detection method for LLMs, BehaviorBench specifically targets the alignment between an LLM's generated behavior and the personality to which it has been shaped. Our approach uniquely integrates (1) psychological theory, (2) empirically validated behaviors, and (3) ecologically plausible scenarios. As a result, we enable a more behaviorally faithful assessment of LLM personality. Table 1 presents example items.

### 3.3 Evaluation and Metrics

We assess personality shaping by first converting each dataset instance into a standardized prompt. For traits other than Neuroticism, we use the following template:

> `<Personality>` \n In the following situation, will you `<behavior>`? Answer 'Yes' or 'No' without explanation.\n Situation: `<Scenario>`

For Neuroticism, where behaviors manifest as cognitive biases assessed via forced-choice questions, we adopt a two-option format:

> `<Personality>`\n In the situation, which one do you prefer? No need to explain the reason.\n Situation: `<Scenario>` `<choice[0]>` or `<choice[1]>`?

In both cases, `<Personality>` is the instructed prompt via different personality shaping methods, `<behavior>` is the target action, and `<Scenario>` is the contextual vignette from $S_b$. For Neuroticism, `<choice[0]>` and `<choice[1]>` correspond randomly to the bias-consistent and bias-inconsistent options, with selection of the bias-consistent option being mapped to a "Yes" response.

We then measure how an injected personality trait $\tau$ influences the model's behavioral responses compared to a neutral baseline (i.e., no explicit personality instruction). For each trait $\tau$, we define a set of theoretically associated behaviors $B_\tau$. For each behavior $b \in B_\tau$, we generate a collection of contextual scenarios $S_b$. The model is prompted with a scenario $s \in S_b$ and asked whether it would perform behavior $b$. Its response is binarized as:

$$\mathrm{LLM}(b, s) = \begin{cases} 1 & \text{if the response is "Yes",} \\ 0 & \text{if the response is "No".} \end{cases}$$

We compute the **behavioral endorsement rate** ($r$) for trait $\tau$ under a given trait setting as:

$$r_\tau = \frac{1}{|B_\tau|} \sum_{b \in B_\tau} \left( \frac{1}{|S_b|} \sum_{s \in S_b} \mathrm{LLM}(b, s) \right) \times 100\%. \quad (1)$$

This metric reflects the model's average propensity to endorse behaviors aligned with trait $\tau$. The larger $r$ is, the better it aligns with a high level of trait. The smaller $r$ is, the better it aligns with the lower trait level.

For each personality trait $\tau$, let $d_\tau \in \{+1, -1\}$ denote the theoretical direction of the instructed setting relative to the no shaping baseline: (1) $d_\tau = +1$ if the instruction implies a higher level of $\tau$ (e.g., "high extraversion"); (2) $d_\tau = -1$ if it implies a lower level (e.g., "low neuroticism"). Let $r_\tau^{\text{inst}}$ be the behavioral endorsement rate under the given personality instruction, and $r_\tau^{\text{nosh}}$ the rate under no shaping conditions. We define the **Personality Influence Score (PIS)** as:

$$\text{PIS}_\tau = d_\tau \cdot \left(r_\tau^{\text{inst}} - r_\tau^{\text{nosh}}\right). \quad (2)$$

A positive $\text{PIS}_\tau$ indicates that the model's behavior shifts in the theoretically expected direction. The overall PIS is computed as the average across all evaluated traits: $\text{PIS} = \frac{1}{N}\sum_{i=1}^{N} \text{PIS}_{\tau_i}$, where $N$ is the total number of traits.

# 4 Experiments

## 4.1 Experimental Settings

Subsequent experiments in this paper are all on BehaviorBench and based on the following configuration.

**Personality shaping methods.** We evaluate one baseline and two widely used prompt-based personality shaping strategies, all grounded in the Big Five framework: (1) No shaping baseline. A generic prompt containing no personality descriptors. (2) Quantified prompts. Numerical trait scores with scale interpretation (e.g., "Your openness score is 5: ..."), adapted from Liu *et al.* [2024]. (3) Descriptive prompts. Natural-language trait descriptions for high/low levels (e.g., "You are an open person with a vivid imagination ..."), following Jiang *et al.* [2023]. The fine-tuning method such as BIG5-CHAT is excluded because it is trained on the dialogue-based dataset, creating a domain mismatch with the scenario-based tasks of BehaviorBench, which can distort general behavioral reasoning and confuse interpretation.

**LLMs.** We evaluate commonly used LLMs, including the closed-source models GPT-4o and Gemini-2.5, as well as the open-source models Qwen3 and DeepSeek-v3.2. Unless otherwise specified, all subsequent analytical experiments adopt the Qwen3-14B model, and all LLMs use the official default inference settings.

**Compared personality assessment method.** We use the Big Five Inventory (BFI) from PsychoBench [Huang *et al.*, 2023b], which is the most commonly adopted self-report questionnaire in LLM personality evaluation [Hartley *et al.*, 2025; Wang *et al.*, 2024; Zou *et al.*, 2024], for comparison to examine how the incorporation of behaviors affects the model's responses.

**Metrics.** We use the behavioral endorsement rate ($r$) and the Personality Influence Score (PIS) to measure the LLMs' behavioral tendency and the effectiveness of personality shaping methods, respectively. See Sec. 3.3.

## 4.2 Main Results

Under the baseline shown in Table 2, BFI scores and BehaviorBench endorsement rates ($r$) often show consistent directional trends, which demonstrates the rationality of BehaviorBench. Qwen3-14B achieves $r = 100\%$ in Agreeableness yet only scores 3.2 on BFI, while DeepSeek-v3.2 shows high $r$ for both Agreeableness (82.0%) and Neuroticism (81.0%) despite moderate BFI scores. These findings suggest that BFI reflects a model's constructed self-narrative, potentially shaped by linguistic priors or social desirability, whereas BehaviorBench captures its actual behavioral tendencies in context. The partial directional agreement confirms that personality signals exist in LLMs, but the substantial magnitude gaps demonstrate that self-report alone is insufficient for reliable personality assessment.

| LLM | O. | | C. | | E. | | A. | | N. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $r$ | BFI | $r$ | BFI | $r$ | BFI | $r$ | BFI | $r$ | BFI |
| **DeepSeek-v3.2** | 30.0 | 4.3 | 59.0 | 3.9 | 70.0 | 3.1 | 82.0 | 3.8 | 81.0 | 3.2 |
| **GPT-4o** | 59.0 | 4.3 | 63.0 | 4.4 | 54.6 | 3.6 | 94.0 | 4.2 | 58.0 | 1.6 |
| **Gemini-2.5** | 42.0 | 4.2 | 48.0 | 4.8 | 40.9 | 3.6 | 80.0 | 4.8 | 65.0 | 1.0 |
| **Qwen3-235B** | 83.0 | 4.5 | 55.0 | 4.8 | 68.2 | 3.5 | 99.0 | 4.7 | 61.0 | 1.0 |
| **Qwen3-32B** | 74.0 | 4.3 | 66.0 | 4.3 | 70.9 | 4.0 | 97.0 | 4.4 | 47.0 | 2.2 |
| **Qwen3-14B** | 83.0 | 3.5 | 41.0 | 3.7 | 60.9 | 3.0 | 100 | 3.2 | 49.0 | 3.0 |

Table 2: Baseline personality of LLMs. For each Big Five trait: Openness (O.), Conscientiousness (C.), Extraversion (E.), Agreeableness (A.), and Neuroticism (N.), we report the behavioral endorsement rate ($r$) and the self-reported BFI score (in 1-5). Gray-shaded cells indicate consistent trends.

Table 3 shows behavioral endorsement rates ($r$) and Personality Influence Scores (PIS) for multiple LLMs under high- and low-trait prompts using quantified and descriptive prompts. Overall, all models exhibit strong plasticity and consistency in both $r$ and PIS under high- and low-trait instructions. This effect is particularly pronounced under quantified prompts, where models demonstrate clearer behavioral shifts and stronger alignment with target personality traits. The consistent and directional responses suggest that LLMs possess a robust understanding of the Big Five personality traits and can effectively translate numerical trait specifications into corresponding behavioral tendencies. Performance varies across models, with GPT-4o leading under quantified prompts. Neuroticism is least affected by personality shaping, likely due to challenges in modeling emotional volatility. Notably, models start with high baseline Agreeableness (see Table 2) and modulate it easily, except DeepSeek-v3.2 with low PIS, suggesting that the strong responsiveness may be linked to the intensity of human preference alignment during RLHF training. In contrast, descriptive high-trait prompts often yield non-significant or negative PIS, likely because natural language descriptions struggle to unambiguously convey personality traits, leading models to misinterpret or overgeneralize the intended behavioral cues.

## 4.3 BehaviorBench vs. BFI

Table 4 presents a personality shaping comparison between self-reported BFI scores and behaviorally observed traits from BehaviorBench ($\hat{r}$, rescaled to the 1–5 scale). With quantified prompts, directional alignment is observed for Conscientiousness, Extraversion, and particularly Agreeableness. However, a pronounced discrepancy emerges for Neuroticism: despite extremely low BFI scores (1.0 under low-

| LLM | Openness | | Conscientiousness | | Extraversion | | Agreeableness | | Neuroticism | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Quantified Prompts** (High & Low-trait $r^{\triangle\,\mathrm{PIS}}$) | | | | | | | | | | | | |
| **Qwen3-235B** | $\mathbf{80.0}^{\triangle\,-3.0}$ | $49.0^{\triangle\,+34.0}$ | $88.0^{\triangle\,+33.0}$ | $23.0^{\triangle\,+32.0}$ | $94.5^{\triangle\,+26.4}$ | $27.3^{\triangle\,+40.9}$ | $\mathbf{100.0}^{\triangle\,+1.0}$ | $14.0^{\triangle\,+85.0}$ | $64.0^{\triangle\,+3.0}$ | $62.0^{\triangle\,-1.0}$ | $85.3^{\triangle\,+12.1}$ | $35.1^{\triangle\,+38.2}$ |
| **DeepSeek-v3.2** | $34.0^{\triangle\,+4.0}$ | $14.0^{\triangle\,+16.0}$ | $81.0^{\triangle\,+22.0}$ | $64.0^{\triangle\,-5.0}$ | $75.5^{\triangle\,+5.5}$ | $32.7^{\triangle\,+37.3}$ | $98.0^{\triangle\,+16.0}$ | $58.0^{\triangle\,+24.0}$ | $53.0^{\triangle\,-28.0}$ | $52.0^{\triangle\,+29.0}$ | $68.3^{\triangle\,+3.9}$ | $44.1^{\triangle\,+20.3}$ |
| **GPT-4o** | $76.0^{\triangle\,+17.0}$ | $18.0^{\triangle\,+41.0}$ | $\mathbf{89.0}^{\triangle\,+26.0}$ | $\mathbf{6.0}^{\triangle\,+57.0}$ | $\mathbf{100.0}^{\triangle\,+45.5}$ | $11.8^{\triangle\,+42.7}$ | $\mathbf{100.0}^{\triangle\,+6.0}$ | $\mathbf{13.0}^{\triangle\,+81.0}$ | $71.0^{\triangle\,+13.0}$ | $\mathbf{49.0}^{\triangle\,+9.0}$ | $\mathbf{87.2}^{\triangle\,+\mathbf{21.5}}$ | $\mathbf{19.6}^{\triangle\,+\mathbf{46.1}}$ |
| **Gemini-2.5** | $51.0^{\triangle\,+9.0}$ | $\mathbf{10.0}^{\triangle\,+32.0}$ | $80.0^{\triangle\,+32.0}$ | $13.0^{\triangle\,+35.0}$ | $48.2^{\triangle\,+7.3}$ | $\mathbf{7.3}^{\triangle\,+33.6}$ | $\mathbf{100.0}^{\triangle\,+20.0}$ | $14.0^{\triangle\,+66.0}$ | $\mathbf{78.0}^{\triangle\,+13.0}$ | $59.0^{\triangle\,+6.0}$ | $71.4^{\triangle\,+16.3}$ | $20.7^{\triangle\,+34.5}$ |
| **Average** | $60.3^{\triangle\,+6.8}$ | $22.8^{\triangle\,+30.8}$ | $84.5^{\triangle\,+28.3}$ | $26.5^{\triangle\,+29.8}$ | $79.6^{\triangle\,+21.2}$ | $19.8^{\triangle\,+38.6}$ | $99.5^{\triangle\,+10.8}$ | $24.8^{\triangle\,+64.0}$ | $66.5^{\triangle\,+0.3}$ | $55.5^{\triangle\,+10.8}$ | $78.1^{\triangle\,+13.5}$ | $29.9^{\triangle\,+31.8}$ |
| **Descriptive Prompts** (High & Low-trait $r^{\triangle\,\mathrm{PIS}}$) | | | | | | | | | | | | |
| **Qwen3-235B** | $\mathbf{89.0}^{\triangle\,+6.0}$ | $29.0^{\triangle\,+54.0}$ | $77.0^{\triangle\,+22.0}$ | $45.0^{\triangle\,+10.0}$ | $\mathbf{33.6}^{\triangle\,-34.5}$ | $20.0^{\triangle\,+48.2}$ | $\mathbf{100.0}^{\triangle\,+1.0}$ | $26.0^{\triangle\,+73.0}$ | $\mathbf{67.0}^{\triangle\,+6.0}$ | $65.0^{\triangle\,-4.0}$ | $\mathbf{73.3}^{\triangle\,+\mathbf{0.1}}$ | $37.0^{\triangle\,+36.2}$ |
| **DeepSeek-v3.2** | $39.0^{\triangle\,+9.0}$ | $26.0^{\triangle\,+4.0}$ | $57.0^{\triangle\,-2.0}$ | $47.0^{\triangle\,+12.0}$ | $24.6^{\triangle\,-45.5}$ | $10.0^{\triangle\,+60.0}$ | $79.0^{\triangle\,-3.0}$ | $19.0^{\triangle\,+63.0}$ | $60.0^{\triangle\,-21.0}$ | $57.0^{\triangle\,+24.0}$ | $51.9^{\triangle\,-12.5}$ | $31.8^{\triangle\,+32.6}$ |
| **GPT-4o** | $61.0^{\triangle\,+2.0}$ | $18.0^{\triangle\,+41.0}$ | $\mathbf{84.0}^{\triangle\,+21.0}$ | $52.0^{\triangle\,+11.0}$ | $17.3^{\triangle\,-37.3}$ | $2.7^{\triangle\,+51.8}$ | $97.0^{\triangle\,+3.0}$ | $5.0^{\triangle\,+89.0}$ | $50.0^{\triangle\,-8.0}$ | $\mathbf{54.0}^{\triangle\,+4.0}$ | $61.9^{\triangle\,-3.9}$ | $26.3^{\triangle\,+\mathbf{39.4}}$ |
| **Gemini-2.5** | $14.0^{\triangle\,-28.0}$ | $\mathbf{7.0}^{\triangle\,+35.0}$ | $72.0^{\triangle\,+24.0}$ | $\mathbf{41.0}^{\triangle\,+7.0}$ | $1.8^{\triangle\,-39.1}$ | $\mathbf{0.9}^{\triangle\,+40.0}$ | $83.0^{\triangle\,+3.0}$ | $\mathbf{1.0}^{\triangle\,+79.0}$ | $65.0^{\triangle\,+0.0}$ | $69.0^{\triangle\,-4.0}$ | $47.2^{\triangle\,-8.0}$ | $\mathbf{23.8}^{\triangle\,+31.4}$ |
| **Average** | $50.8^{\triangle\,-2.8}$ | $20.0^{\triangle\,+33.5}$ | $72.5^{\triangle\,+16.3}$ | $46.3^{\triangle\,+10.0}$ | $19.3^{\triangle\,-39.1}$ | $8.4^{\triangle\,+50.0}$ | $89.8^{\triangle\,+1.0}$ | $12.8^{\triangle\,+76.0}$ | $60.5^{\triangle\,-5.8}$ | $61.3^{\triangle\,+5.0}$ | $58.6^{\triangle\,-6.1}$ | $29.7^{\triangle\,+31.9}$ |

Table 3: Behavioral endorsement rates ($r$) under high/low-trait prompts, annotated with Personality Influence Scores ($r^{\triangle\,\mathrm{PIS}}$). Higher $r$ under high-trait (or lower under low-trait) prompts indicates better personality alignment. PIS magnitude shows shift size from baseline; sign ($+/-$) indicates expected or opposite direction. Best results are **bolded**.

| Metric | O. | | C. | | E. | | A. | | N. | |
|---|---|---|---|---|---|---|---|---|---|---|
| Quantified Prompts (High and Low trait) | | | | | | | | | | |
| BFI | 4.3 | 1.4 | 3.7 | 1.0 | 3.2 | 1.0 | 3.4 | 1.9 | 3.8 | 1.0 |
| $\hat{r}$ | 3.3 | 2.6 | 3.1 | 1.2 | 3.2 | 1.7 | 5.0 | 2.4 | 3.2 | 3.1 |
| Descriptive Prompts (High and Low trait) | | | | | | | | | | |
| BFI | 3.1 | 1.8 | 3.8 | 2.4 | 2.5 | 1.5 | 3.8 | 2.3 | 4.2 | 3.0 |
| $\hat{r}$ | 3.0 | 2.3 | 3.5 | 2.2 | 1.4 | 1.3 | 5.0 | 1.7 | 2.4 | 1.3 |

Table 4: Comparison of BFI and BehaviorBench scores. The $r$ values are linearly rescaled to the 1–5 range as $\hat{r} = 1 + 0.04r$ to align with the BFI scale for direct comparison. For both metrics, scores above 3 reflect high trait expression; those below 3, low expression.

| Trait | O. | C. | E. | A. | N. |
|---|---|---|---|---|---|
| **w/o** | 83.0 | 41.0 | 60.9 | 100 | 49.0 |
| **1** | $41.0^{\triangle\,+42.0}$ | $5.0^{\triangle\,+36.0}$ | $16.4^{\triangle\,+44.5}$ | $35.0^{\triangle\,+65.0}$ | $52.0^{\triangle\,-3.0}$ |
| **2** | $33.0^{\triangle\,+50.0}$ | $6.0^{\triangle\,+35.0}$ | $12.7^{\triangle\,+48.2}$ | $26.0^{\triangle\,+74.0}$ | $54.0^{\triangle\,-5.0}$ |
| **3** | $42.0^{\triangle\,+41.0}$ | $11.0^{\triangle\,+30.0}$ | $20.9^{\triangle\,+40.0}$ | $85.0^{\triangle\,+15.0}$ | $56.0^{\triangle\,-7.0}$ |
| **4** | $53.0^{\triangle\,-30.0}$ | $42.0^{\triangle\,+1.0}$ | $30.9^{\triangle\,-30.0}$ | $100.0^{\triangle\,+0.0}$ | $54.0^{\triangle\,+5.0}$ |
| **5** | $58.0^{\triangle\,-25.0}$ | $53.0^{\triangle\,+12.0}$ | $55.5^{\triangle\,-5.5}$ | $100.0^{\triangle\,+0.0}$ | $54.0^{\triangle\,+5.0}$ |

Table 5: Behavioral endorsement rates $r$ and PIS ($r^{\triangle\,\mathrm{PIS}}$) under single-trait quantified prompts across 5 fine-grained trait levels.

trait), $\hat{r}$ remains near the neutral midpoint (3.1), indicating a failure of behavioral expression to reflect reported emotional stability. Agreeableness is most consistent across prompts. These findings highlight a gap between self-reports and behavior in LLMs, showing that behavioral assessment is essential, particularly for traits like Neuroticism, as self-reports alone can greatly overstate personality consistency.

## 4.4 Analysis of Fine-grained Quantified Prompts

As shown in Table 5, we evaluate single-trait quantified prompts across five fine-grained intensity levels (1 to 5), measuring $r$ and PIS relative to the baseline. Overall, the model exhibits a clear relationship for most traits: as the specified trait level increases, $r$ generally rises monotonically, indicating sensitivity to prompt granularity. Conscientiousness (C.) and Extraversion (E.) show strong controllability. From baselines (41.0 and 60.9, respectively), low-level prompts substantially suppress behavioral expression, while high-level prompts (4–5) effectively elevate $r$ back toward or above baseline. Both Agreeableness (A.) and Openness (O.) yield high baseline representations, and low-intensity prompts can effectively reduce $r$ for both traits, demonstrating that even strongly default behaviors can be suppressed. However, high-level prompts fully restore $r$ to the ceiling state for Agreeableness, whereas those for Openness only achieve a partial recovery of $r$. In stark contrast, Neuroticism (N.) remains largely unaffected across all levels, with $r$ fluctuating narrowly around the baseline and PIS values close to zero, confirming its resistance to quantified personality prompting.

| Mode | O. | | C. | | E. | | A. | | N. | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quantified Prompts (High & Low-trait) | | | | | | | | | | | | |
| Single | 58.0 | **41.0** | **53.0** | **5.0** | 55.5 | **16.4** | **100.0** | **35.0** | **54.0** | **52.0** | 64.1 | 29.9 |
| Multi | **61.4** | 56.9 | 49.8 | 21.6 | **63.0** | 24.8 | 98.6 | 66.1 | 53.4 | 54.2 | 65.2 | 44.7 |
| Descriptive Prompts (High & Low-trait) | | | | | | | | | | | | |
| Single | 50.0 | 32.0 | **62.0** | 30.0 | 10.9 | **7.3** | **99.0** | **17.0** | 35.0 | **37.0** | 51.4 | 24.7 |
| Multi | **53.2** | **29.5** | 61.9 | **22.0** | **42.3** | 18.9 | 98.6 | 46.2 | **43.2** | 44.0 | 59.8 | 32.1 |

Table 6: Behavioral endorsement rates ($r$) under single-trait and multi-trait prompting settings. For each trait, the better result between single- and multi-trait conditions is **bolded**.

## 4.5 Analysis of Multi-trait Personality Shaping

To investigate how the presence of other traits influences the expression of a target trait, we conduct a comprehensive evaluation across all possible settings of the Big Five traits. Specifically, we consider all $2^5 = 32$ combinations, where each trait is independently set to either a high or low level. For each target trait, we compute the average $r$ between the intended and generated personality scores, marginalizing over all $2^4 = 16$ settings of the remaining four traits. The results in Table 6 show that, overall, the behavioral expression of a target trait remains largely consistent between single- and multi-trait prompting, particularly for high-trait conditions, indicating that LLMs can maintain trait-specific behaviors even in the presence of compound personality signals. However, under low-trait conditions, the multi-trait setting consistently yields smaller deviations from the midpoint, suggesting that concurrent traits suppress the deterministic influence of any

| Think | O. | | C. | | E. | | A. | | N. | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quantified Prompts (High & Low-trait) | | | | | | | | | | | | |
| *w/* | **72.0** | **21.0** | **99.0** | **0.0** | **89.1** | **6.4** | 99.0 | **0.0** | **61.0** | **49.0** | **84.0** | **15.3** |
| *w/o* | 58.0 | 41.0 | 53.0 | 5.0 | 55.5 | 16.4 | **100.0** | 35.0 | 54.0 | 52.0 | 64.1 | 29.9 |
| Descriptive Prompts (High & Low-trait) | | | | | | | | | | | | |
| *w/* | **52.0** | **24.0** | **79.0** | 41.0 | 7.3 | **0.0** | 97.0 | **5.0** | **40.0** | 38.0 | **55.1** | **21.6** |
| *w/o* | 50.0 | 32.0 | 62.0 | **30.0** | **10.9** | 7.3 | **99.0** | 17.0 | 35.0 | **37.0** | 51.4 | 24.7 |

Table 7: Behavioral endorsement rates ($r$) on Qwen3-14B with and without Think inference mode.

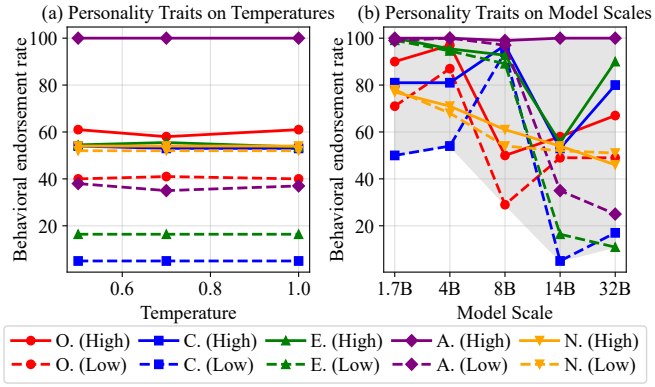


Figure 3: Behavioral endorsement rates ($r$) under quantified prompts for (a) different temperature settings (0.5, 0.7, and 1.0) on Qwen3-14B, and (b) across various Qwen3 model sizes (1.7B – 32B) with temperature = 0.7.

single trait on behavior. The reason is that multi-dimensional personality descriptions inherently constrain the dominance of any single negative trait.

## 4.6 Factors Affecting Personality Controllability

**Think Mode.** We evaluate the impact of enabling Think mode on personality controllability. As Table 7 shows, across both quantified and descriptive prompts, activating Think mode improves the LLM's adherence to target traits, while low-trait prompts achieve stronger suppression. This demonstrates that Think mode enhances the LLM's ability to interpret and faithfully execute personality instructions.

**Temperature.** To further investigate the impact of different sampling strategies on personality controllability, we conduct experiments using three temperature settings: 0.5, 0.7 (default), and 1.0, under quantified prompts. As Figure 3 (a) shows, temperature has little effect on personality expression.

**Model Scales.** We also explore how model size impacts personality controllability by evaluating the Qwen3 series models with sizes 1.7B, 4B, 8B, 14B, and 32B. The experimental results are presented in Figure 3 (b). LLMs exhibit greater separation between high- and low-trait behavioral responses with larger scales, indicating heightened sensitivity to personality prompts.

**Robustness.** As shown in Table 8, we report variances from experiments with three independent runs and three prompt templates under quantified prompts. The average variance between repeated runs is 2.1, confirming high stability. Most traits maintain low variance across templates, except low agreeableness, indicating that LLMs' inherent high agreeableness makes this trait harder to modulate via prompts.

| O. | | C. | | E. | | A. | | N. | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Three runs (High & Low-trait) | | | | | | | | | | |
| 0.3 | 1.0 | 1.0 | 0.3 | 2.5 | 0.3 | 0.0 | 4.0 | 1.3 | 10.3 | 2.1 |
| Three prompt templates (High & Low-trait) | | | | | | | | | | |
| 2.3 | 30.3 | 9.0 | 6.3 | 13.5 | 5.2 | 0.0 | 142.3 | 2.3 | 4.3 | 21.6 |

Table 8: Variance of behavioral endorsement rates ($r$) across three independent repeated experimental runs and three prompt templates using quantified prompts method on Qwen3-14B.

| Character | Metric | Profile | O. | C. | E. | A. | N. |
|---|---|---|---|---|---|---|---|
| James Bond | BFI | - | 3.7 | 5.0 | 3.3 | 3.3 | 1.3 |
| | $\hat{r}$ | w/ | 2.6 | 4.0 | 1.4 | 4.6 | 3.2 |
| | $\hat{r}$ | w/o | 3.3 | 3.8 | 3.3 | 5.0 | 3.2 |
| Rorschach | BFI | - | 1.7 | 4.7 | 1.7 | 1.7 | 4.0 |
| | $\hat{r}$ | w/ | 2.2 | 2.6 | 1.0 | 1.1 | 3.2 |
| | $\hat{r}$ | w/o | 3.2 | 2.3 | 1.9 | 4.2 | 3.1 |
| Gaston | BFI | - | 1.7 | 2.0 | 4.7 | 1.0 | 4.7 |
| | $\hat{r}$ | w/ | 3.0 | 2.4 | 1.1 | 1.8 | 3.7 |
| | $\hat{r}$ | w/o | 3.9 | 1.9 | 3.1 | 4.8 | 3.4 |

Table 9: Comparison of target personality (BFI, 1–5 scale) and model behavior ($\hat{r}$, rescaled to 1–5) under quantified prompts on Qwen3-14B. BFI values and profiles are from Wang *et al.* [2024].

## 4.7 Impact of Individual Profiles

We examine whether adding a narrative character profile alters the behavioral expression of personality under fixed quantified trait prompts. We test three well-known characters: James Bond from *Tomorrow Never Dies*, Rorschach from *Watchmen*, and Gaston from Disney's *Beauty and the Beast*, each with established Big Five scores from [Wang *et al.*, 2024] (shown in BFI rows of Table 9). Notably, for roles with a typically low Agreeableness (e.g., Rorschach, Gaston), profiles significantly reduce $r$ relative to profile-free prompts. This indicates that character profiles play an active role in guiding personality-behavior alignment in LLMs, not just providing surface-level narrative framing.

# 5 Conclusion

This work presents BehaviorBench, a psychologically grounded benchmark that evaluates LLM personality through realistic behavioral responses rather than self-report scales. We demonstrate that quantified prompting enables more precise personality control than descriptive methods, and reveal a critical mismatch between self-reported and behavioral traits, especially for Neuroticism, highlighting limitations in current personality shaping paradigms. Crucially, we show that explicit reasoning significantly enhances personality fidelity, offering a practical pathway toward building LLMs with stable and controllable personalities. Future work can extend this framework to richer interaction scenarios and alternative psychological models beyond the Big Five.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62472252), the Natural Science Foundation of Shandong Province, China (Grant No. ZR2025QC1567), and the Research Projects of Quancheng Laboratory, China (Grant Nos. QCL20250105 and QCL20250204).

## References

[Al Otaibi, 2012] Sameera Moharib B Al Otaibi. The relationship between cognitive dissonance and the big-5 factors model of the personality and the academic achievement in a sample of female students at the university of umm al-qura. *Education*, 132(3):607–624, 2012.

[Ashton *et al.*, 1998] Michael C Ashton, Sampo V Paunonen, Edward Helmes, and Douglas N Jackson. Kin altruism, reciprocal altruism, and the big five personality factors. *Evolution and Human Behavior*, 19(4):243–255, 1998.

[Belhekar, 2017] Vivek M Belhekar. Cognitive and non-cognitive determinants of heuristics of judgment and decision-making: General ability and personality traits. *Journal of the Indian Academy of Applied Psychology*, 43(1):75, 2017.

[Besta *et al.*, 2025] Maciej Besta, Shriram Chandran, Robert Gerstenberger, Mathis Lindner, Marcin Chrapek, Sebastian Hermann Martschat, Taraneh Ghandi, Patrick Iff, Hubert Niewiadomski, Piotr Nyczyk, et al. Psychologically enhanced ai agents. *arXiv preprint arXiv:2509.04343*, 2025.

[Blais and Weber, 2006] Ann-Renée Blais and Elke U Weber. A domain-specific risk-taking (dospert) scale for adultpopulations. *Judgment and Decision making*, 1(1):33–47, 2006.

[Breil *et al.*, 2019] Simon M Breil, Katharina Geukes, Robert E Wilson, Steffen Nestler, Simine Vazire, and Mitja D Back. Zooming into real-life extraversion–how personality and situation shape sociability in social interactions. *Collabra: Psychology*, 5(1):7, 2019.

[Briggs, 1976] Katharine C Briggs. *Myers-Briggs type indicator*. Consulting Psychologists Press Palo Alto, CA, 1976.

[Caron and Srivastava, 2023] Graham Caron and Shashank Srivastava. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, 2023.

[Cobb-Clark and Schurer, 2012] Deborah A Cobb-Clark and Stefanie Schurer. The stability of big-five personality traits. *Economics Letters*, 115(1):11–15, 2012.

[Deng *et al.*, 2025] Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao yang, Xin Zhao, and Ji-Rong Wen. Neuron based personality trait induction in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[DeYoung *et al.*, 2002] Colin G DeYoung, Jordan B Peterson, and Daniel M Higgins. Higher-order factors of the big five predict conformity: Are there neuroses of health? *Personality and Individual differences*, 33(4):533–552, 2002.

[Guo *et al.*, 2025] Yaoqi Guo, Zhenpeng Chen, Jie M Zhang, Yang Liu, and Yun Ma. Personality-guided code generation using large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1068–1080, 2025.

[Han *et al.*, 2025] Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *arXiv preprint arXiv:2509.03730*, 2025.

[Hartley *et al.*, 2025] John Hartley, Conor Brian Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, and Raad Khraishi. How personality traits shape llm risk-taking behaviour. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21068–21092, 2025.

[Hilbig, 2008] Benjamin E Hilbig. Individual differences in fast-and-frugal decision making: Neuroticism and the recognition heuristic. *Journal of Research in Personality*, 42(6):1641–1645, 2008.

[Huang *et al.*, 2023a] Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*, pages 1–8, 2023.

[Huang *et al.*, 2023b] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*, 2023.

[Jensen-Campbell *et al.*, 2007] Lauri A Jensen-Campbell, Jennifer M Knack, Amy M Waldrip, and Shaun D Campbell. Do big five personality traits associated with self-control influence the regulation of anger and aggression? *Journal of research in personality*, 41(2):403–424, 2007.

[Jiang *et al.*, 2023] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

[John *et al.*, 1991] Oliver P John, Eileen M Donahue, and Robert L Kentle. Big five inventory. *Journal of personality and social psychology*, 1991.

[Kahneman and Tversky, 2013] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[Karra *et al.*, 2022] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. Estimating the personal-

ity of white-box language models. *arXiv preprint arXiv:2204.12000*, 2022.

[Kipman *et al.*, 2021] U Kipman, M Weib, S Bartholdy, G Schiepek, and W Aichom. Personality and risk taking. *Austin J Clin Case Rep*, 8(9):1233, 2021.

[Lee *et al.*, 2025] Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, 2025.

[Li *et al.*, 2025] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, 2025.

[Liu *et al.*, 2024] Jianzhi Liu, Hexiang Gu, Tianyu Zheng, Liuyu Xiang, Huijia Wu, Jie Fu, and Zhaofeng He. Dynamic generation of personalities with large language models. *arXiv preprint arXiv:2404.07084*, 2024.

[Mao *et al.*, 2018] Tianxin Mao, Weigang Pan, Yingying Zhu, Jian Yang, Qiaoling Dong, and Guofu Zhou. Self-control mediates the relationship between personality trait and impulsivity. *Personality and Individual Differences*, 129:70–75, 2018.

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

[McElroy and Dowd, 2007] Todd McElroy and Keith Dowd. Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and Decision making*, 2(1):48–53, 2007.

[Meng and Leung, 2021] Keira Shuyang Meng and Louis Leung. Factors influencing tiktok engagement behaviors in china: An examination of gratifications sought, narcissism, and the big five personality traits. *Telecommunications Policy*, 45(7):102172, 2021.

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

[Neijenhuis and Kupiainen, 2023] Ellen V Neijenhuis and Melanie S Gava Kupiainen. Understanding the relationship between the personality trait neuroticism and swift blame in an organizational setting. 2023.

[Oda *et al.*, 2014] Ryo Oda, Wataru Machii, Shinpei Takagi, Yuta Kato, Mia Takeda, Toko Kiyonari, Yasuyuki Fukukawa, and Kai Hiraishi. Personality and altruism in daily life. *Personality and Individual Differences*, 56:206–209, 2014.

[Paulhus and Williams, 2002] Delroy L Paulhus and Kevin M Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563, 2002.

[Roberts and Yoon, 2022] Brent W Roberts and Hee J Yoon. Personality psychology. *Annual review of psychology*, 73(1):489–516, 2022.

[Rushton *et al.*, 1981] J Philippe Rushton, Roland D Chrisjohn, and G Cynthia Fekken. The altruistic personality and the self-report altruism scale. *Personality and individual differences*, 2(4):293–302, 1981.

[Tangney *et al.*, 2018] June P Tangney, Angie Luzio Boone, and Roy F Baumeister. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. In *Self-regulation and self-control*, pages 173–212. Routledge, 2018.

[Tok, 2011] Serdar Tok. The big five personality traits and risky sport participation. *Social Behavior and Personality: an international journal*, 39(8):1105–1111, 2011.

[Tu *et al.*, 2023] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*, 2023.

[Tversky and Kahneman, 1974] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

[Wang *et al.*, 2024] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, 2024.

[West *et al.*, 2022] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 4602–4625, 2022.

[Zhang *et al.*, 2025] Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, et al. Omnicharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. *arXiv preprint arXiv:2505.20277*, 2025.

[Zou *et al.*, 2024] Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. Can llm" self-report"?: Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots. *arXiv preprint arXiv:2412.00207*, 2024.