

Designing AI-Agents with Personalities: A Psychometric Approach

Muhua Huang^{1 2}, Xijuan Zhang⁴, Christopher Soto⁵, and James Evans^{2 3}

¹Stanford University

²University of Chicago Knowledge Lab

³Chicago Center for Computational Social Science, Chicago

⁴York University

⁵Colby College

Author Note

This article has been conditionally accepted for publication in the *Personality Science* journal following peer review. This is a pre-copyedited, author-produced version of the article.

Correspondence concerning this article should be addressed to Muhua Huang, Stanford Graduate School of Business. Contact: muhua@stanford.edu.

Abstract

We introduce a methodology for assigning quantifiable and psychometrically validated personalities to AI-Agents using the Big Five framework. Across three studies, we evaluate its feasibility and limitations. In Study 1, we show that large language models (LLMs) capture semantic similarities among Big Five measures, providing a basis for personality assignment. In Study 2, we create AI-Agents using prompts designed based on the Big Five Inventory-2 (BFI-2) in different format, and find that AI-Agents powered by new models align more closely with human responses on the Mini-Markers test, although the finer pattern of results (e.g., factor loading patterns) were sometimes inconsistent. In Study 3, we validate our AI-Agents on risk-taking and moral dilemma vignettes, finding that models prompted with the BFI-2-Expanded format most closely reproduce human personality-decision associations, while safety-aligned models generally inflate ‘moral’ ratings. Overall, our results show that AI-Agents align with humans in correlations between input Big Five traits and output responses and may serve as useful tools for preliminary research. Nevertheless, discrepancies in finer response patterns indicate that AI-Agents cannot (yet) fully substitute for human participants in precision or high-stakes projects.

Keywords: Artificial Intelligence, AI Agents, Large Language Model, Simulation, Big Five Personalities, Psychometrics

Designing AI-Agents with Personalities: A Psychometric Approach

Introduction

The emergence of large language models (LLMs) has revolutionized our approach to simulating human-like behaviors and communication. LLMs are increasingly deployed across diverse research fields to mimic human behaviors (Xi et al., 2023; Xu et al., 2024). In psychology, LLMs are used to study cognitive processes, measure personality, and understand psychological constructs (Binz et al., 2025; Hagendorff et al., 2023; Jiang et al., 2024; Rathje et al., 2024). Sociologists employ these models to explore social bias and behavior (Lucy & Bamman, 2021; Park et al., 2023; H. Zhang et al., 2025), while economists and political scientists utilize them to analyze economic processes and political leanings (Bang et al., 2024; Hartmann et al., 2023). These broad applications underscore the growing interconnection between artificial intelligence and the social and behavioral sciences.

Traditional human subjects research methods, while standard in social and behavioral studies, face significant challenges including ethical constraints, logistical hurdles, and resource limitations (Demszky et al., 2023; Salganik, 2019). Recent advancements in LLMs offer valuable complementary tools to address these challenges (Agnew et al., 2024; Demszky et al., 2023). Researchers can use LLMs to create AI-Agents, which can mimic responses from human participants, reducing logistical and financial burdens (S. Wang et al., 2021). These AI-Agents operate continuously, simulate diverse demographic responses, and can be deployed in scenarios that might pose ethical risks to human participants (Aher et al., 2023; Argyle et al., 2023; Bai et al., 2022).

The integration of AI-Agents in social and behavioral science research serves multiple practical purposes, complementing rather than replacing human participants. They can be used for pilot testing studies, allowing researchers to refine experimental designs and identify potential issues before investing in full-scale human participant studies. Additionally, AI-Agents can independently replicate findings from human-subjects data, enhancing the robustness and generalizability of research

outcomes (Kozlowski et al., 2024; H. Zhang et al., 2025). This approach provides a complementary method for validation and exploration, contributing to the overall rigor and efficiency of social and behavioral science research. By offering a novel approach to corroborate and extend findings obtained through traditional human participant studies, AI-Agents represent an advancement in social science research methodology, enhancing inclusivity and efficiency while preserving the essential depth of human-centered understanding critical in the field.

Problems with Previous Approaches for Assigning Personas to AI-Agents

Among the myriad traits that can be incorporated into AI-Agent personas, personalities stand out as among the most intuitive characteristics to simulate. Personalities encapsulate a spectrum of human behaviors and tendencies crucial for the prediction of life outcomes ranging from academic and career to health and socioeconomic outcomes (Soldz & Vaillant, 1999; Soto, 2019; Stewart et al., 2022) to a multitude of interaction-based phenomena (Dang & Tapus, 2014; Furnham & Heaven, 1999). Despite this potential and the capacity of AI-Agents to advance personality research, existing methodologies for implementing personality in AI-Agents remain limited, typically falling into three distinct approaches.

First, personas are assigned with simple personality adjectives in prompts. For example, Jiang et al. (2024) assigned personalities to an AI-Agent by telling it “You are a character who is introverted, antagonistic, conscientious, emotionally stable, and open to experience” This type of approach holds a binary view, with either high-low or presence-absence of a trait, which contradicts established empirical evidence that personality traits exist on continuous spectra (Asendorpf, 2006; Marcus et al., 2006; Zrari & Sakale, 2024).

Second, personas may be assigned through demographic descriptions and personal preferences (Bai et al., 2025; Park et al., 2023; Serapio-García et al., 2023). For instance, Park et al. (2023) used narrative descriptions such as “John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people.” This approach relies heavily on stereotypical information extrapolated from the LLM’s

training corpora, where names and occupations may trigger implicit assumptions (e.g., associating "John Lin" with Asian ethnicity and pharmaceutical expertise). This approach, however, poses difficulties in evaluating the full scope of stereotypical implications (i.e., potential confounds) and fails to provide the granularity and precision necessary for social and behavioral science studies, particularly when the degree of trait expression is crucial for creating AI-Agents with specific personality profiles.

Third, personas may be embedded at the parameter level through methods like fine-tuning or direct weight editing. In one approach, researchers fine-tune an LLM on an individual-specific or group-specific text corpus, producing a distinct model persona for each dataset (Liu et al., 2024; Tan et al., 2025). Alternatively, knowledge editing techniques modify the model's internal weights or latent representations to encode desired traits. For example, recent work identifies "persona vectors", which are the specific directions in a model's activation space associated with character traits (Chen et al., 2025; Kim et al., 2025). Applying such a vector inside the model can induce a target personality (e.g. making the model more sycophantic) without full retraining. While these parameter-level interventions alter behavior more fundamentally than in-context prompts, they still lack fine-grained control over how traits manifest. Moreover, locating and adjusting the correct parameters or latent directions for a given persona is technically challenging. The process often involves complex optimization and substantial computational resources, demanding deep expertise and making it less accessible to researchers without specialized backgrounds or institutional support.

These approaches offer a superficial semblance of personality but suffer from critical limitations. Relying on stereotypes underlying traits (e.g., naive) or roles (e.g., policing officer) does not provide the precise control necessary for rigorous social and behavioral research and does not reflect the continuity and complexity inherent in human personalities (Cummings & Sanders, 2019).

Psychometric Approach for Assigning LLM-Agents Persona

In this article, we propose creating AI-Agents with personalities through a psychometric approach by leveraging Big Five Personality theory. Historically, Big Five

personality theory was developed based on the *Lexical Hypothesis*, which states that personality characteristics fundamental to humans have become a part of human language, and the most important characteristics are encoded by a single word (Caprara & Cervone, 2000). Based on the Lexical Hypothesis, scholars selected personality-related words from the English dictionary (Allport & Odbert, 1936), refined the list of words (W. T. Norman, 1963; W. Norman, 1967; Wiggins, 1979), had participants rate themselves against those descriptors, applied factor analysis to reduce the dimensionality of the data, and eventually identified the *Big Five Personality traits* (Fiske, 1949; Goldberg, 1990; McCrae, 1994; Peabody & Goldberg, 1989): Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N).

Over the past decades, many psychometric scales¹ measuring the Big Five traits have been developed, notably the Big Five Inventory (BFI; John et al., 1991), Big Five Inventory-2 (BFI2; Soto & John, 2017), Mini-Markers (Saucier, 1994), the Big Five Aspects Scale (BFAS; DeYoung et al., 2007) and the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1989). Despite having different items, these psychological measurements demonstrated high reliability, convergent validity and predictive validity in samples from diverse backgrounds (McCrae, 2009; McCrae & Costa Jr, 1997; Poortinga et al., 2002). Specifically, these psychological measurements of the Big Five can significantly predict life outcomes, such as educational achievement, socioeconomic status, health, interpersonal relationships with peers and family, and more, indicating that personality traits can indeed forecast a range of individual, interpersonal and societal behaviors effectively (Ozer & Benet-Martinez, 2006; Roberts et al., 2007; Soto, 2019).

Given that the Big Five traits and LLMs are both developed based on natural language (English), using the Big Five framework to create AI-Agents with personality is theoretically intuitive and coherent. Therefore, in this article, we propose a new prompt engineering method that leverages items and response options from a popular Big Five scale to assign personalities to AI-Agents. Given the high reliability and

¹ The terms “scale” and “test” are used interchangeably to refer to psychometric instruments for assessing personality traits.

validity of the Big Five and the rigorous research on the Big Five theory over the decades, AI-Agents created with Big Five prompts will have personalities that are more realistic, nuanced, and fine-grained, reflecting the distribution of personality types and variability in a given population.

Bringing psychometric principles to design and evaluate has practical benefits (X. Wang et al., 2023; Ye et al., 2025). Integrating well-established psychometric tests (i.e., high reliability and validity) into the design of AI-Agents could ensure that the personality traits assigned to AI-Agents are stable and reflective of their programmed characteristics, enhancing the credibility and robustness of findings in applications where the consistency of AI-Agents performance is critical. In addition, as psychometric tests are designed to discern different levels of expression related to psychological traits, we can reverse engineer that variability into AI-Agents to create diversity, which is necessary for simulating real-world settings. Additionally, free from external confounds, this approach allows for precise and fine-grained manipulation over the psychological constructs of interest, allowing nuanced understanding and refinement of AI-Agents' personalities for tailored applications.

In summary, leveraging the Big Five personality test to design prompts for personality assignment of AI-Agents presents an opportunity to move beyond stereotypes and create AI-Agents with psychometrically valid traits. These traits can be quantified and controlled, are realistic and predictive of real human behavior, and are thus better suited for large-scale deployment in psychology and social science research.

Objectives of the Present Research

The main goal of the current research is to develop and validate AI-Agents with personalities created using psychometrically sound Big Five personality measures. To achieve this goal, we conducted a series of three interconnected studies. In Study 1, we aim to establish a foundational understanding of personality constructs and measurements using a modern LLM technique called embedding, which converts natural language into numeric vectors based on semantic relatedness between words, phrases and sentences. Using this technique, we examine how personality-related constructs are

semantically represented in LLMs.

In Study 2, we examine how to use prompt engineering to create AI-Agents with personality using a popular and validated personality measure called the Big-Five Inventory-2 (BFI-2). As a comparison, we also included baseline conditions where we created AI-Agents using simple adjectives. To validate these AI-agents, we prompted AI-agents to complete a criterion measure (i.e., the Mini-Marker test) and compared their responses to human participants' responses with matching personalities. In other words, we assess whether AI-Agents' responses align with human participants' responses across different AI-Agent conditions. We hypothesize that AI agents created using the BFI-2 will capture more nuances in personalities and thus show greater alignment with human responses than those created using simple adjectives.

In Study 3, we further validate the AI-Agents by prompting them to respond to real-life moral and risk-taking vignettes and examining the extent to which their responses align with human participants' responses. We again hypothesize that AI agents created using the BFI-2 will align better with human responses than those created using simple adjectives.

Collectively, these three studies form a comprehensive investigation into the creation, validation, and application of AI-Agents with psychometrically sound personality traits. We hope to provide researchers with a powerful new tool for conducting personality and social behavior research at scale, while maintaining high standards of validity and reliability.

Data Availability Statement

All data, research materials, and analysis code have been made publicly available through GitHub (<https://github.com/muhua-h/Psychometrics4AI>). Study 3 was preregistered on the Open Science Framework prior to data collection (<https://osf.io/4t9bf>) and received ethical approval from the Ethics Review Board at [Institution Name] (ID: e2024-221). The sample size, exclusion criteria, and analysis plan for Study 3 were specified in the preregistration.

Study 1: Semantic Representation of Personality Constructs in LLM

Study 1 aimed to establish a foundational understanding of the semantic nuances inherent in personality tests and constructs using embedding. Because the behavior of AI-Agents is ultimately driven by embeddings, this initial analysis serves to validate the approach and set the stage for subsequent studies.

Methods

Our analysis incorporated content from widely recognized Big Five tests, such as the BFI (John et al., 1991), BFI-2 (Soto & John, 2017), Mini-Markers (Saucier, 1994), BFAS (DeYoung et al., 2007) and NEO-FFI (Costa & McCrae, 1989). We extracted domain-specific content (i.e., test items) from these tests and processed it through OpenAI’s advanced text-embedding model (`text-embedding-3-large`). Text embeddings are commonly used in natural language processing; they allow researchers to represent text data (such as words, phrases, sentences, or even entire documents) in a numeric format (i.e., vector) that quantifies the semantic relationships between words, phrases, or entire documents. In the context of our study, each personality test item is converted into a high-dimensional vector (3072 dimensions for the model we used) that captures its semantic meaning.

To illustrate this concept, consider three items from different Big Five tests:

- (a) “Is outgoing, sociable” (BFI-2, Extraversion)
- (b) “Extraverted” (Mini-Markers, Extraversion)
- (c) “Is original, comes up with new ideas” (BFI-2, Openness)

Items (a) and (b), despite their different wording, are semantically similar and would be represented by vectors close to each other in the high-dimensional space, with a similarity of .65 as measured by *cosine distance*. By contrast, item (c) measures a different construct, although it follows the same wording style as item (a). Item (c)’s similarity to item (a) is .33, and to item (b) .23. This example shows the extent to which embeddings capture convergent and divergent psychological constructs.

Embeddings Techniques

To analyze these embeddings, we employed two techniques, namely, cosine similarity and *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE).

Cosine similarity. This measure quantifies the semantic similarity between two embedded texts, analogous to how correlation coefficients measure the relationship between variables in psychological research. Just as a correlation of 1 indicates a perfect positive relationship, a cosine similarity of 1 indicates identical semantic meaning. In contrast, a cosine similarity of 0 suggests no semantic relationship, similar to a correlation of 0 indicating no linear relationship between variables.

In the context of personality assessment, cosine similarity can be thought of as a measure of construct overlap between items or scales. For instance, high cosine similarity between items from different tests (e.g., “Is outgoing, socioable” from BFI2 and “Extraverted” from Mini-Markers) would suggest they are tapping into the same underlying construct, similar to how high inter-item correlations within a scale indicate internal consistency in classical test theory.

***t*-SNE.** We applied *t*-SNE (Van der Maaten & Hinton, 2008) to item embeddings to create a two-dimensional visualization of their similarity structure. *t*-SNE is a non-linear dimensionality reduction method that maps high-dimensional data to a low-dimensional space while preserving local relationships between data points. In our context, the *t*-SNE provides an intuitive visual map of semantic neighbourhoods among personality test items, which provides a good complement to the cosine similarity analysis. Items from the same Big Five domain tend to cluster together in the plot, and cross-domain proximities become easier to spot visually. In our *t*-SNE visualizations, items that cluster together can be interpreted as measuring similar constructs².

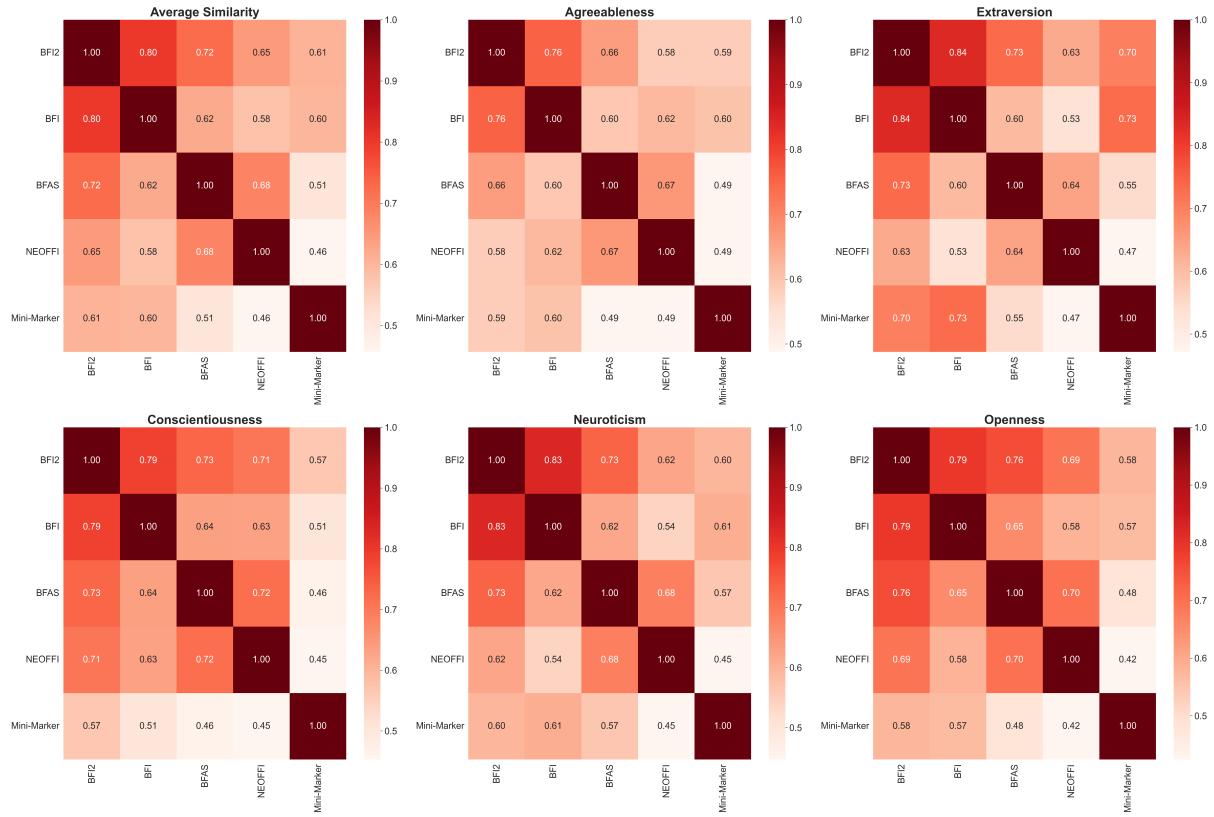
The use of embedding in this study offers several advantages for personality research. By capturing semantic relationships between test items, we can potentially uncover subtle distinctions between various personality assessments that might not be

² Because *t*-SNE is intended for visualization rather than formal measurement modelling, the plots are used here descriptively to illustrate patterns in the embedding space; the axes have no intrinsic psychological meaning.

apparent through traditional psychometric methods. This approach provides a data-driven complement to classical item analysis techniques, potentially informing more refined personality models and enhancing our understanding of the semantic overlap between different personality measures.

Figure 1

Cosine Similarity Between Personality Tests: Overall Average and Domain-Specific Comparisons



Results and Discussion

Figure 1 illustrates the cosine similarity between different personality tests across the Big Five domains. The top-left subplot displays the average similarity across all domains, indicating that most tests generally exhibit moderate to high cosine similarity (above 0.51), except for Mini-Markers and NEO-FFI. The Mini-Markers test consistently shows a relatively lower cosine similarity with other scales across all domains, which may be attributed to its unique design approach: while other tests employ full statements, phrases, or questions for each item (e.g., “Is complex, a deep thinker”), the Mini-Markers test exclusively uses adjectives (e.g., philosophical”). The

remaining subplots, each corresponding to a specific Big Five domain, reveal similar patterns of cosine similarity between tests. This consistency across domains suggests that semantic relationships between different personality tests are relatively stable, regardless of the specific trait being measured. However, varying degrees of similarity observed between different test pairs highlight the nuanced differences in how each instrument operationalizes and measures personality constructs within the Big Five framework.

Figure 2

Two-Dimensional Projection of Big Five Personality Test Domain Embeddings Using t-SNE

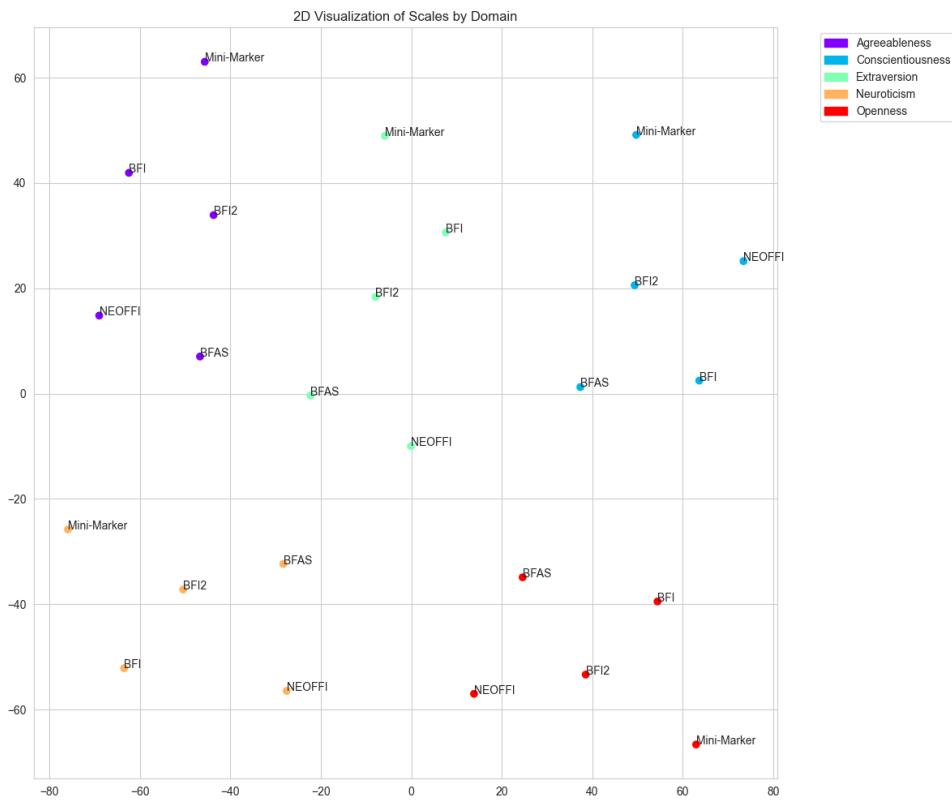


Figure 2 shows a two-dimensional projection of the tests' domain embeddings using t-SNE³, which complements the statistical results by offering an intuitive view of how items from different tests are positioned relative to each other in the embedding

³ Although t-SNE is typically applied to larger datasets, it can still be used with smaller datasets for illustrative purposes. Here, it is employed solely for visualization, not for statistical inference. The axes for t-SNE plot are usually not labeled, because the units of the axes are not directly related to any measurement from the original data. Instead, the axes represent a space constructed to best represent similarities and differences between data points.

space. Each point represents a personality test item, and items from the same Big Five domain tend to cluster together, mirroring within-domain similarities shown in Figure 1. This visualization reveals distinct clustering patterns for each Big Five personality domain, with the spatial arrangement of these clusters providing insight into the semantic relationships between personality constructs. For instance, items assessing Agreeableness are concentrated in the upper left quadrant, while those measuring Openness are predominantly situated in the lower right quadrant. The remaining domains (i.e., Extraversion, Conscientiousness, and Neuroticism) form separate, distinguishable clusters within the projection space.

These results suggest that different personality tests tap into highly similar and consistent constructs, despite variations in item wording and test structure. This finding serves as the foundation for subsequent studies, which further examine Agents' understanding of the underlying semantic associations during the personality assignment, adaptation, and reflection process.

Study 2: Creating and Validating AI-Agents with Personalities

Considering the findings of Study 1, which indicated a relatively low semantic similarity between the BFI-2 and the Mini-Markers test, in Study 2, we aim to create AI-Agents using the BFI-2 and then validate the personality assignments using the Mini-Markers test. The rationale is that if this method supports two fairly distinct psychometric tests, it should generalize to other tests with greater semantic similarity.

Study 2 consists of two parts: Study 2a and Study 2b. In Study 2a, we discuss how to use prompt engineering to create AI-agents with personalities and examine the alignment between AI-agents' responses on the Mini-Markers test responses and those of human participants collected in a previous study. In Study 2b, we demonstrate how to create and validate AI-Agents representative of human participants' sample data using only summary statistics from a previous study, offering an efficient way to create AI-agents without collecting new human data.

Study 2a

Method

Prompt Engineering. We included four prompt conditions for creating AI-Agents with personalities (see Table 1). Two conditions involved using the BFI-2 in either Likert or Expanded format, which we refer to as *BFI-2-Likert* and *BFI-2-Expanded*, respectively. The BFI-2-Likert is the original BFI-2 by Soto and John (2017), in which participants indicated their agreement with statements measuring the Big Five traits. The BFI-2-Likert, however, is susceptible to response bias and careless responding; therefore, X. Zhang et al. (2025) developed the BFI-2-Expanded version, replacing the agree-disagree response options with complete sentences. For example, for the Expanded format, the “strongly agree” option for the item “Is outgoing, sociable” is replaced with “I am very outgoing, sociable”. For prompt engineering, the Expanded format produces prompts that sound more natural and straightforward. As shown in Table 1, the BFI-2-Likert prompt requires an additional instruction for interpreting numeric numbers indicating agreement levels, whereas the BFI-2-Expanded prompt simply describes the trait levels in complete sentences.

We also included two baseline prompt conditions (Simple-Binary and Elaborated-Binary), designed to mimic the approach of previous studies that used simple adjectives to create AI-Agent personas Jiang et al. (2024), Park et al. (2023), and Serapio-García et al. (2023). In the Simple-Binary condition, prompts were binary Big Five statements indicating whether the AI-Agent was high or low on each trait. The Elaborated-Binary condition added descriptors such as “outgoing” and “compassionate” to further elaborate on the traits. Compared with these baseline conditions, we aim to show that the BFI-2-Likert or BFI-2-Expanded condition create AI-Agents with more nuanced personalities and aligns better with human responses.

Data. To examine alignment between AI-Agents responses with those from human participants, we repurposed data collected by Soto and John (2017) wherein participants ($N = 438$) responded to multiple Big Five tests, including: (1) BFI-2: a sophisticated 60-item Likert test designed to capture the comprehensive hierarchical

Table 1*Example Prompts for Creating AI-Agents With Personalities*

Simple-Binary	Elaborated-Binary
### Context### You are participating in a personality psychology study. You have been assigned personality traits.	### Context### You are participating in a personality psychology study. You have been assigned personality traits.
### Your Personality ### You are high in Extraversion. You are high in Agreeableness. You are high in Conscientiousness. You are low in Neuroticism. You are high in Openness.	### Your Personality ### You are high in Extraversion. You are outgoing, sociable, assertive, and energetic. You are high in Agreeableness. You are compassionate, cooperative, trusting, and kind to others. You are high in Conscientiousness. You are organized, responsible, hardworking, and reliable. You are low in Openness. You prefer familiar routines, practical approaches, and conventional ideas.
BFI-2-Likert	BFI-2-Expanded
### Context ### You are participating in a personality psychology study. You have been assigned personality traits.	### Context ### You are participating in a personality psychology study. You have been assigned personality traits.
### Instruction### Each number below indicates the extent to which you agree or disagree with that statement. 1 means 'Disagree Strongly', 2 means 'Disagree', 3 means 'Neutral', and 4 means 'Agree', 5 means 'Agree Strongly'.	### Your Personality ### I am very outgoing, sociable. I am very compassionate, almost always soft-hearted. I am fairly organized. I am somewhat relaxed, handle stress somewhat well. I have some artistic interests. I am quite assertive. I am very respectful almost always treat others with respect. I am often lazy. I stay very optimistic after experiencing a setback. I am curious about few things. I often feel excited or eager. I rarely find fault with others. I am very dependable, steady. I am fairly moody often have up and down mood swings. I am fairly inventive, often find clever ways to do things. I am rarely quiet. I feel a great deal of sympathy for others. I am not particularly systematic rarely keep things in order. I am sometimes tense. I am very much fascinated by art, music or literature. I am fairly submissive, often act as a follower. I rarely start arguments with others. I have a fair amount of difficulty getting started on tasks. I feel very secure, comfortable with self. I typically seek out intellectual, philosophical discussions. I am somewhat more active than other people. I have a very forgiving nature. I am often careless. I am fairly emotionally stable quite hard to upset. I have little creativity. I am often shy, introverted. I am very helpful and unselfish with others. I sometimes keep things neat and tidy. I worry quite a lot. I value art and beauty quite a bit. I find it fairly easy to influence people. I am sometimes rude to others. I am fairly efficient, get things done fairly quickly. I almost never feel sad. I am not particularly complex rarely a deep thinker. I am almost always full of energy. I am quite trusting of others' intentions. I am fairly reliable, can usually be counted on. I usually keep my emotions under control. I have a bit of difficulty imagining things. I am very talkative. I am very warm and caring. I rarely leave a mess usually clean up. I often feel anxious or afraid. I think poetry and plays are fairly interesting. I strongly prefer to take charge. I am very polite and courteous to others. I am fairly persistent, usually work until the task is finished. I sometimes feel depressed and blue. I have little interest in abstract ideas. I show a lot of enthusiasm. I almost always assume the best about people. I often behave irresponsibly. I am not at all temperamental almost never get emotional. I am not at all original almost never come up with new ideas.

structure of personality, where each domain has three facets and (2) Mini-Markers: a straightforward, 40-item test consisting of phenotypic Big Five descriptive adjectives (Saucier, 1994). Participants' BFI2 responses will serve as the training (input) data, while their Mini-Markers responses (output) will be used for validation. The human data (Soto & John, 2017) will serve as a reference for our comparison.

Procedure. Using the four types of prompts in Table 1, we created AI-Agents across five popular LLMs, including 1) GPT-3.5-Turbo (01-25), 2) GPT-4-Turbo (04-09), 3) GPT-4o (2024-11-20), 4) DeepSeek-V3 (2024-12-26), and 5) Llama-3.3-70B-Instruct (2024-12-06). For each LLM and each prompt condition, we initiated 438 AI-Agents, each corresponding to a human participant's data collected from Soto and John (2017). More specifically, for the BFI-2-Likert and BFI-2-Expanded conditions, AI-Agents were created by inputting prompts that matched the BFI-2 scores of the corresponding human participants. For the Baseline-Simple and Baseline-Elaborated conditions, participants' trait scores were dichotomized as high (average score > 2.5) or low (average score < 2.5) for each of the Big Five traits, and AI-Agents were then created using these dichotomized scores. Across all conditions, the LLM temperature was set to the default value of 1.0 to balance coherence and diversity, and other sampling parameters (e.g., top-p, frequency, and presence penalties) were kept at their defaults.

After creating AI-Agents for each LLM and prompt condition, we prompted them to complete the Mini-Markers test, which are 40 adjectives describing the Big Five traits (Saucier, 1994). The full prompt for this task is provided in the Appendix A. In this prompt, AI-Agents were instructed to consider the Mini-Marker items carefully based on their assigned personality (see Appendix).

Data Analysis

Comparing Mini-Marker Scores Between AI-Agents and Humans. For each LLM, prompt condition, and personality trait, we conducted correlational analyses, paired-sample t -tests, and Kolmogorov-Smirnov (KS) tests to compare the Mini-Markers responses of AI-agents with those of humans who had matching

underlying BFI-2 personalities. Correlational analyses examined the linear relationships between AI-agent and human responses, paired-sample *t*-tests compared their mean Mini-Markers scores, and KS tests compared their score distributions. For all tests, Bonferroni corrections are used to control for family-wise Type I error rate.

Correlational Analyses Between BFI-2 and Mini-Markers. In addition, for each LLM and personality trait, we computed the correlation between the Big Five and the Mini-Markers. Specifically, for the Likert and Expanded prompt conditions, we correlated the Mini-Marker output scores with the BFI-2-Likert and -Expanded input scores (via prompts), respectively.

For the Baseline-Simple and Baseline-Elaborate conditions, we computed two types of correlations. First, we correlate the original BFI-2-Likert scores (from which the binary inputs were derived) with the Mini-Markers output scores. Second, we correlate the binary Big Five input scores with the Mini-Markers scores. We add the second set of correlations because LLMs only received the binary inputs in those two conditions without access to the original BFI-2-Likert scores. By comparing these correlations from AI-Agents to those from humans (i.e., human participants' correlation between BFI-2 and Mini-markers), we can determine whether AI-agents can help researchers recover correlations between variables in research.

Factor Analyses. Finally, we conducted Confirmatory Factor Analyses (CFA)⁴ using the *lavaan* package in R (Rosseel, 2012). For each LLM, prompt condition, and personality trait, we fitted a one-factor model and compute the one-factor reliability coefficient (a.k.a., Omega). To evaluate the fit of the two CFA models, we used the chi-square test of fit and three approximate fit indices with common cutoff points: 1) the comparative fit index (CFI), with a value above .90 indicating a reasonable fit and above .95 indicating a very good fit; 2) the root mean square error of approximation

⁴ CFA is a widely used psychometric method for testing whether a set of questionnaire items reflects an expected structure of psychological traits(Kline, 2014). For example, in the Big Five model, items about being “outgoing” or “sociable” should cluster together as indicators of Extraversion. CFA formally tests such hypotheses by estimating how strongly each item loads onto a latent factor (e.g., Extraversion) and by providing indices of overall model fit (e.g., CFI, RMSEA, SRMR). Good fit means the data are consistent with the hypothesized trait structure, while poor fit suggests the responses may not capture the intended personality dimensions. In our study, CFA allows us to evaluate whether AI-Agent responses reproduce the same underlying Big Five structure observed in human data.

(RMSEA), with a value of less than .08 indicating reasonable fit and less than .05 indicating very good fit; and 3) the standardized root mean square residual (SRMR), with a value of less than .08 indicating reasonable fit and less than .05 indicating very good fit.

Results and Discussion

Comparing Mini-Marker Scores Between AI-Agents and Humans.

Table 2 shows the mean differences in average Mini-Marker scores between AI-Agents and human responses across prompt and LLM conditions. Table 3 presents the KS test statistics, which measure the maximum absolute differences between the distributions of AI-Agents and human responses. Figure 3 shows graphs visualizing the mean and distributional differences for the Conscientiousness average scores (see Supplementary Materials for the graphs for the other four traits).

As shown in Tables 2 and 3, the means and distributions differed significantly between AI-Agents and human responses in the Simple- and Elaborated-Binary prompt conditions. As illustrated in Figure 3, except for GPT-3.5, AI-Agents' Mini-Markers scores in these conditions were essentially binary, with most responses clustering at the very high end and some at the very low end, resulting in higher mean scores than those observed in humans. In other words, for GPT-4, GPT-4o, Llama, and DeepSeek, when the assigned personality prompts consisted of simple binary adjectives, the models interpreted the underlying assigned personality as binary, either very high or very low on a given trait. By contrast, for humans, regardless of whether traits were measured dichotomously or continuously, underlying distributions were continuous and typically normally distributed. Interestingly, GPT-3.5, the older and smaller model, tended to generate more moderate and variable results even under the Simple-Binary and Elaborated-Binary conditions. This suggests that as the LLMs grow bigger and their capability improves, the models become more sensitive to personality steering prompts, making simple personality descriptions less suitable to simulate human populations.

For the BFI-2-Likert and -Expanded prompt conditions, although most LLMs still produced AI-Agents with mean scores significantly different from those of humans

(see Table 2), many generated distributions that closely resembled those of humans. As shown in Table 3 and Figure 3, the score distributions of GPT-4, GPT-4o, Llama, and DeepSeek were particularly similar to human distributions, especially under the BFI-2-Likert condition. Interestingly, GPT-3.5 once again diverged from this pattern: consistent with its behavior in the Simple- and Elaborated-Binary conditions, its AI-Agent responses were more centered around the mid-point (i.e., around 5) and exhibited a more bell-shaped distribution.

Taken together, these results indicate that the bigger and newer LLMs (GPT-4, GPT-4o, Llama, and DeepSeek) tend to reproduce the response outputs more consistent with the prompt inputs. Specifically, when the prompt presented the Big Five traits as either high or low (i.e., in the Simple- and Elaborated-Binary conditions), AI-Agent responses also fell at the extremes; when the prompt presented varying levels of the traits, responses formed a continuum. Consequently, newer LLMs showed poor alignment with human responses under the Simple- and Elaborated-Binary conditions but much better alignment under the BFI-2-Likert and -Expanded conditions. By contrast, the older GPT-3.5 consistently generated more moderate responses across prompt conditions, resulting in mediocre alignment across conditions.

Correlational Analyses Between BFI-2 and Mini-Markers. Table 4, for the BFI-2-Likert and -Expanded conditions, shows domain-level Pearson correlations between BFI-2 input scores and Mini-Markers output scores. For the Simple- and Elaborated-Binary conditions, Table 4 includes two sets of correlations: (1) correlations between the binary Big Five input scores and Mini-Markers output scores (right-hand values), and (2) correlations between the original BFI-2 Likert scores (from which binary inputs were derived) and Mini-Markers output scores (left-hand values). As a reference, the first row of Table 4 shows the correlations for human participants between their BFI-2-Likert and Mini-Markers scores (Human: O=.75, C=.84, E=.88, A=.80, N=.74; Avg=.80; $N = 438$).

Across all conditions, the correlation coefficients in Table 4 were significantly different from zero, indicating that Big Five input and Mini-Marker output were

Table 2*Mean Differences between AI-Agents and Human Responses on the Mini-Markers*

Condition	LLM	O	C	E	A	N
Simple-Binary	GPT-3.5	-0.58	-0.04	-0.79	-0.48	0.94
Simple-Binary	GPT-4	-1.42	-1.34	-1.78	-1.55	0.03
Simple-Binary	GPT-4o	-1.88	-1.38	-1.94	-1.55	-0.93
Simple-Binary	Llama	-1.19	-1.12	-1.91	-1.55	0.77
Simple-Binary	DeepSeek	-1.54	-1.32	-1.84	-1.82	0.35
Elaborated-Binary	GPT-3.5	-0.59	-0.11	-0.79	-0.52	0.95
Elaborated-Binary	GPT-4	-1.42	-1.34	-1.78	-1.55	-0.08
Elaborated-Binary	GPT-4o	-1.88	-1.36	-1.94	-1.55	-0.91
Elaborated-Binary	Llama	-1.22	-1.13	-1.91	-1.57	0.77
Elaborated-Binary	DeepSeek	-1.58	-1.35	-1.81	-1.77	0.26
BFI-2-Likert	GPT-3.5	1.71	0.84	0.43	0.86	0.95
BFI-2-Likert	GPT-4	1.22	0.24	0.47	0.29	0.26
BFI-2-Likert	GPT-4o	0.20	0.15	0.35	0.15	0.20
BFI-2-Likert	Llama	0.04	0.14	-0.01	-0.12	0.33
BFI-2-Likert	DeepSeek	0.40	0.19	0.02	0.30	0.07
BFI-2-Expanded	GPT-3.5	0.44	0.96	0.29	0.29	0.43
BFI-2-Expanded	GPT-4	0.17	0.65	0.22	0.22	0.44
BFI-2-Expanded	GPT-4o	0.02	0.57	0.29	0.53	-0.34
BFI-2-Expanded	Llama	-0.13	0.46	-0.02	0.25	0.29
BFI-2-Expanded	DeepSeek	-0.30	0.39	-0.14	0.20	0.35

Note: Unshaded cells indicate significant paired-sample *t*-tests comparing AI-Agent and human responses, whereas gray-shaded cells indicate non-significant results. Significance levels were adjusted using Bonferroni corrections to control the family-wise Type I error rate.

significantly related across conditions. For the BFI-2-Likert and BFI-2-Expanded conditions, across LLMs and traits, correlations between the BFI-2 input and the Mini-Markers output scores were generally high (Likert Avgs: GPT-3.5=.75 GPT-4=.86, GPT-4o=.91, Llama=.90, DeepSeek=.90; Expanded Avgs: .78, .86, .90, .90, .89).

On the other hand, for Simple- and Elaborated-Binary conditions, across LLMs and traits, right-hand correlations (between binary Big Five input and Mini-Markers output scores) were generally high, but left-hand correlations (between the original BFI-2 and Mini-Markers scores) were substantially lower. For example, under the Simple-Binary for GPT-4, Conscientiousness' left-hand and right-hand correlations were .58 and .94, respectively, Agreeableness were .35 and .91; for GPT-4o, Openness were .45 and .81 and Agreeableness were .42 and .88. This split mirrors the earlier item-score comparison: LLMs tend to produce quasi-binary Mini-Marker outputs under binary

Table 3

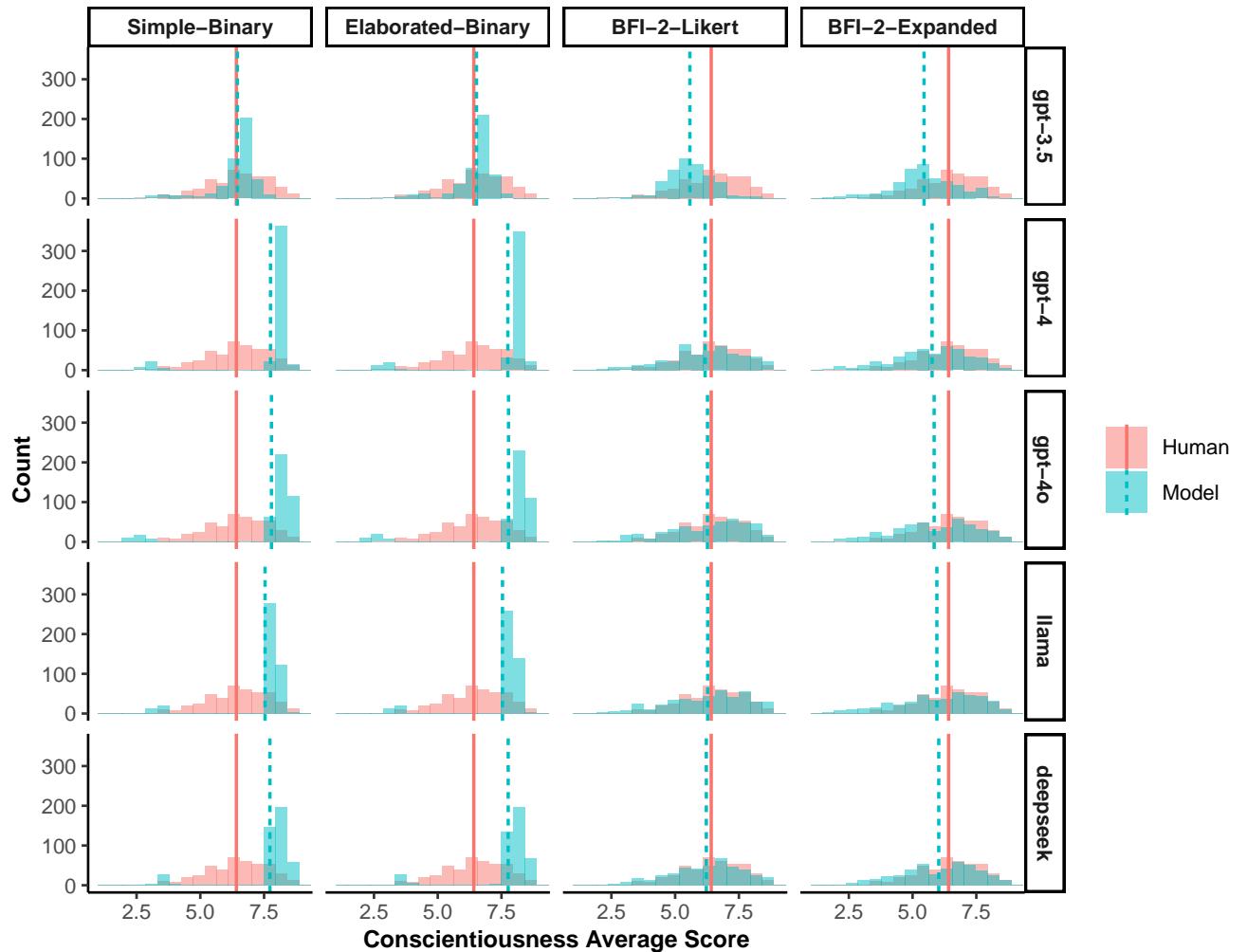
Kolmogorov-Smirnov Test Statistic Comparing AI-Agents and Human Score Distributions for the Mini-Markers

Condition	LLM	O	C	E	A	N
Simple-Binary	GPT-3.5	0.47	0.23	0.56	0.42	0.45
Simple-Binary	GPT-4	0.77	0.78	0.74	0.84	0.25
Simple-Binary	GPT-4o	0.87	0.76	0.76	0.79	0.59
Simple-Binary	Llama	0.77	0.72	0.78	0.85	0.32
Simple-Binary	DeepSeek	0.79	0.76	0.75	0.87	0.27
Elaborated-Binary	GPT-3.5	0.44	0.21	0.55	0.42	0.46
Elaborated-Binary	GPT-4	0.77	0.78	0.75	0.84	0.28
Elaborated-Binary	GPT-4o	0.87	0.77	0.76	0.80	0.59
Elaborated-Binary	Llama	0.78	0.71	0.78	0.85	0.33
Elaborated-Binary	DeepSeek	0.79	0.75	0.75	0.85	0.27
BFI-2-Likert	GPT-3.5	0.62	0.39	0.22	0.44	0.37
BFI-2-Likert	GPT-4	0.42	0.11	0.19	0.14	0.07
BFI-2-Likert	GPT-4o	0.13	0.10	0.20	0.12	0.09
BFI-2-Likert	Llama	0.07	0.09	0.13	0.21	0.11
BFI-2-Likert	DeepSeek	0.19	0.08	0.09	0.15	0.06
BFI-2-Expanded	GPT-3.5	0.13	0.38	0.12	0.19	0.21
BFI-2-Expanded	GPT-4	0.09	0.21	0.16	0.12	0.13
BFI-2-Expanded	GPT-4o	0.07	0.19	0.18	0.20	0.14
BFI-2-Expanded	Llama	0.12	0.14	0.20	0.14	0.10
BFI-2-Expanded	DeepSeek	0.18	0.15	0.22	0.11	0.13

Note: Unshaded cells indicate significant paired-sample *t*-tests comparing AI-Agent and human responses, whereas gray-shaded cells indicate non-significant results. Significance levels were adjusted using Bonferroni corrections to control the family-wise Type I error rate.

prompts, which (1) mismatch continuous human Likert scores, which depresses left-hand correlations, but (2) aligns tightly with the binary inputs, elevating right-hand correlations. The high right-hand values indicate successful tracking of the prompted high/low signal, whereas the low left-hand values remind us that such tracking does *not* necessarily reproduce the underlying continuous human trait covariation. The inflated correlations are a warning for researchers who attempt to approximate human population distribution using binary personality assignment.

Across models, newer and more capable models achieved higher convergence compared to older ones. For example, average correlations for GPT-3.5, GPT-4, and GPT-4o, were .75, .86, .91 for BFI-2-Likert condition, and .78, .86, .90 for BFI-2-Expanded condition. By domain, Conscientiousness and Extraversion are consistently strongest for newer models in richer formats (e.g., Likert C≈.91-.93;

Figure 3*Mini-Markers' Conscientiousness Scores For Human and AI-Agents*

Note. The red line indicates the mean of the human responses. The dotted blue lines indicate the means of AI-Agents across conditions.

$E \approx .92\text{-.94}$), Agreeableness was the main laggard for GPT-3.5 (Likert $A = .67$; Expanded $A = .64$), and Openness showed the widest model spread under Likert (GPT-3.5 = .60 vs Llama = .89), with Expanded partially narrowing gaps for older models.

Regarding human-AI alignment, Table 4 shades in gray the cells that are not statistically different from the human reference (adjusted with Bonferroni-corrected Fisher-z); unshaded cells indicate significant differences. There are few noticeable patterns of results. First, for the BFI-2-Likert and -Expanded conditions, the AI-Agents and humans had very good alignment for the Openness and Extraversion traits and some alignment for the Conscientiousness and Agreeableness traits, with the Expanded format showing better alignment than the Likert format. Second, for Simple- and

Elaborate-Binary conditions, left-handed correlations (between the original BFI-2 and Mini-Markers scores), except for the Neuroticism trait, showed very poor alignment, especially for newer LLMs. On the other hand, right-hand correlations (between the binary Big Five input and Mini-Markers output scores) showed good alignment for the Openness and Extraversion traits but were inflated relative to human correlations for the other traits. These results again underscore that newer LLMs reliably track low/high signals embedded in the prompt inputs, but in doing so they fail to reproduce the subtler, continuous covariation that characterizes human personality traits.

Factor Analyses. Table 8 shows fit measures from the factor analyses. Table 5 shows selected results for factor loadings and reliability coefficients. Full results from factor analyses are provided in the Supplementary Materials.

The most prominent pattern is that for newer LLMs (GPT-4, GPT-4o, Llama, and DeepSeek), the Simple- and Elaborated-Binary conditions yielded very good model fit and extremely high reliability, greatly exceeding those observed for the human condition (see Tables 8 and 5). As shown in Table 5, with GPT-4o factor loadings and reliabilities in the binary conditions were all greater than 0.90, with many reaching 0.99. This pattern of results again reflects stronger consistency between prompt inputs and LLM outputs for the newer models; in the binary conditions, AI-agents tended to produce identical output responses across items, producing extremely high inter-item correlations. For illustration, Table 6 shows the correlation matrix for Neuroticism items in the Elaborated-Binary condition under GPT-4o. The correlations are uniformly high and markedly different from those based on human data, which explains the unrealistically good fit observed for these conditions.

For BFI-2-Likert and -Expanded conditions with the newer LLMs, although fit indices and reliability coefficients were lower than in binary conditions, results were overall more aligned with those from human data; their correlation matrices also more resembled those of humans (see Tables 8 and 5, 6). Despite these relative similarities, notable differences remained. First, AI-Agent in these conditions showed considerably poorer model fit than human data, especially in terms of the CFI and RMSEA values,

Table 4*Correlation between BFI-2 Input and the Mini-Markers Output (Study 2a)*

Condition	Model	O	C	E	A	N	Avg
Human		.75	.84	.88	.80	.74	.80
Simple-Binary	GPT-3.5	.37/.70	.55/.79	.67/.88	.29/.74	.56/.57	.49/.73
Simple-Binary	GPT-4	.42/.79	.58/.94	.67/.89	.35/.91	.76/.88	.56/.88
Simple-Binary	GPT-4o	.45/.81	.62/.91	.69/.88	.42/.88	.77/.90	.59/.88
Simple-Binary	Llama	.42/.82	.58/.92	.67/.88	.33/.88	.78/.87	.56/.87
Simple-Binary	Deepseek	.36/.76	.58/.92	.67/.88	.37/.89	.78/.88	.55/.87
Elaborated-Binary	GPT-3.5	.37/.71	.58/.77	.67/.87	.28/.73	.54/.58	.49/.73
Elaborated-Binary	GPT-4	.42/.77	.58/.93	.68/.89	.37/.92	.77/.88	.56/.88
Elaborated-Binary	GPT-4o	.44/.82	.61/.92	.69/.88	.39/.88	.77/.90	.58/.88
Elaborated-Binary	Llama	.41/.79	.57/.92	.67/.89	.36/.86	.78/.87	.56/.87
Elaborated-Binary	Deepseek	.38/.75	.57/.92	.68/.88	.37/.88	.78/.88	.55/.86
BFI-2-Likert	GPT-3.5	.60	.85	.86	.67	.79	.75
BFI-2-Likert	GPT-4	.69	.91	.92	.90	.89	.86
BFI-2-Likert	GPT-4o	.86	.93	.92	.90	.92	.91
BFI-2-Likert	Llama	.89	.91	.94	.87	.92	.90
BFI-2-Likert	Deepseek	.82	.92	.93	.90	.91	.90
BFI-2-Expanded	GPT-3.5	.78	.79	.86	.64	.83	.78
BFI-2-Expanded	GPT-4	.80	.87	.90	.84	.90	.86
BFI-2-Expanded	GPT-4o	.82	.93	.92	.90	.91	.90
BFI-2-Expanded	Llama	.86	.92	.92	.87	.91	.90
BFI-2-Expanded	Deepseek	.82	.92	.91	.87	.93	.89

Note. For the Simple- and Elaborated-Binary conditions, there are two sets of correlations: (1) the correlations between the binary Big Five input scores and the Mini-Markers output scores (right-hand values), and (2) the correlations between the original BFI-2 Likert scores (from which the binary inputs were derived) and the Mini-Markers output scores (left-hand values). For BFI-2-Likert and -Expanded conditions, the correlations were between the BFI-2 input scores and the Mini-Markers output scores. All the correlation coefficients were significantly different from zero at $\alpha = .05$. Light-gray shading marks values that are *not* statistically different from the Human reference after Bonferroni correction across domains ($\alpha = .05$; two-sided Fisher z test on Fisher-transformed r); unshaded values indicate significant differences.

although, in some Likert conditions, SRMR values were comparable to human data (see Table 8). Second, the pattern of factor loadings differed: items that loaded especially highly for humans did not necessarily load highly for AI-Agents; nonetheless, the overall reliability coefficients were similar between AI-Agents and humans in these conditions (see Table 5).

Finally, for GPT-3.5, Simple- and Elaborated-Binary conditions did not yield as high a reliability as those for newer LLMs. The model fit were generally lower than that of human data across prompt conditions. Except for the Neuroticism subscale, reliabilities for other traits across prompt conditions were comparable to those of human data (see Table 5). The Neuroticism subscale performed poorly with GPT-3.5, yielding reliabilities as low as 0.26 (see Table 5). It seems that the reason for this poor performance is that some items such as “relaxed” and “unenvious” loaded extremely poorly under GPT-3.5’s AI-Agents. Overall, GPT-3.5 performed worse than newer LLMs.

Study 2b

In Study 2b, we present a parametric approach for developing AI-Agents using sample statistics derived from existing personality data. This method involves extracting key parameters from empirical data, simulating item responses based on these parameters, and then assigning these simulated responses to AI-Agents. By doing so, we aim to provide an efficient alternative or precursor to traditional empirical data collection, facilitating the creation of diverse sets of Agents while maintaining psychometric validity.

Table 5*Selected Results for Factor Loadings and Reliability in Study 2a*

Domain	Item	Human	Simple Binary	Elaborated Binary	BFI-2 Likert	BFI-2 Expanded
GPT-3.5						
Neuroticism	Envious	.84	.94	.94	.99	.97
	Fretful	.52	.24	.20	.26	.20
	Jealous	.84	.95	.96	.99	.96
	Moody	.58	.22	.23	.35	.24
	Relaxed	.36	.12	.17	-.09	.00
	Temperamental	.63	.27	.25	.36	.28
	Touchy	.52	.28	.35	.57	.36
	Unenvious	.64	.27	.26	.33	.54
	Reliability	.81	.26	.27	.51	.29
Openness	Complex	.37	.65	.70	.79	.65
	Creative	.51	.94	.93	.83	.38
	Deep	.87	.63	.71	.88	.70
	Imaginative	.82	.94	.95	.85	.44
	Intellectual	.51	.88	.91	.72	.96
	Philosophical	.45	.74	.79	.67	.94
	Uncreative	.72	.32	.37	.10	.32
	Unintellectual	.36	.39	.44	.04	.45
	Reliability	.76	.88	.92	.84	.72
GPT-4o						
Neuroticism	Envious	.84	.98	.98	.69	.94
	Fretful	.52	.99	.99	.69	.64
	Jealous	.84	.98	.98	.71	.94
	Moody	.58	.99	.99	.84	.65
	Relaxed	.36	.98	.98	.61	.61
	Temperamental	.63	.99	.99	.91	.69
	Touchy	.52	.99	.99	.95	.79
	Unenvious	.64	.98	.98	.68	.94
	Reliability	.81	.99	.99	.89	.82
Openness	Complex	.37	.94	.94	.72	.54
	Creative	.51	.97	.98	.60	.35
	Deep	.87	.92	.93	.74	.57
	Imaginative	.82	.98	.99	.62	.34
	Intellectual	.51	.95	.95	.96	.99
	Philosophical	.45	.95	.95	.93	.98
	Uncreative	.72	.93	.93	.57	.39
	Unintellectual	.36	.91	.90	.85	.94
	Reliability	.76	.98	.98	.87	.70

Note. Reliability is computed based on one-factor model reliability (a.k.a., Omega coefficient).

Table 6*Correlation Matrices Between Neuroticism Items In Selected Conditions*

	Envious	Fretful	Jealous	Moody	Temper- amental	Touchy	Relaxed	Un- envious
Human								
Envious	1.00							
Fretful	0.41	1.00						
Jealous	0.76	0.40	1.00					
Moody	0.39	0.44	0.42	1.00				
Temperamental	0.46	0.38	0.45	0.64	1.00			
Touchy	0.39	0.22	0.40	0.45	0.59	1.00		
Relaxed	0.24	0.42	0.23	0.33	0.30	0.20	1.00	
Unenvious	0.60	0.25	0.57	0.31	0.28	0.22	0.22	1.00
Elaborated-Binary with GPT-4o								
Envious	1.00							
Fretful	0.98	1.00						
Jealous	0.99	0.98	1.00					
Moody	0.97	0.99	0.97	1.00				
Temperamental	0.98	0.99	0.98	0.99	1.00			
Touchy	0.98	0.99	0.98	0.99	0.99	1.00		
Relaxed	0.96	0.98	0.96	0.97	0.98	0.98	1.00	
Unenvious	0.98	0.97	0.98	0.97	0.97	0.98	0.96	1.00
BFI-2-Likert with GPT-4o								
Envious	1.00							
Fretful	0.47	1.00						
Jealous	0.89	0.49	1.00					
Moody	0.54	0.57	0.56	1.00				
Temperamental	0.50	0.58	0.54	0.80	1.00			
Touchy	0.61	0.66	0.64	0.79	0.91	1.00		
Relaxed	0.40	0.63	0.41	0.55	0.52	0.56	1.00	
Unenvious	0.83	0.49	0.80	0.57	0.52	0.59	0.45	1.00

Methods and Data Analyses

We generated synthetic BFI-2 data based on sample statistics obtained from Soto and John (2017) empirical sample. To do so, we first extracted key statistics from each personality domain. Specifically, within each domain, we extracted (1) the means and standard deviations for each of the three facets, (2) the 3×3 correlation matrix among the three facets, and (3) average correlation among the four items within each facet (See Supplementary Materials for Sample Code).

Using these sample statistics, we then simulated $n = 200$ participants' underlying Big Five scores under the following assumptions: (1) personality domains were uncorrelated with each other (i.e., orthogonal domains), (2) within each domain, facets were correlated with each other, (3) within each facet, the items were normally distributed.

For the BFI-2-Likert and BFI-2-Expanded prompt conditions, the simulated scores were discretized into five categories corresponding to the BFI-2 response scale. These values were then used to generate Likert- or Expanded-formatted prompts for the LLMs, producing AI-Agents with personalities (see Table 1). For the Simple- and Elaborated-Binary conditions, scores were dichotomized and translated into prompts specifying whether the AI-Agents should be high or low on each trait (see Table 1). The statistical simulation script is available in the Supplementary Material.

After creating AI-Agents with personalities, we prompted the AI-Agents to complete the Mini-Markers test just like in Study 2a. With AI-Agents' responses on the Mini-Markers test, we conducted analyses parallel to some of the analyses in Study 2a: 1) correlational analyses between the simulated BFI-2 input scores and the Mini-Markers scores, 2) Omega reliability coefficients, and 3) CFA.

Results and Discussion

Correlational Analysis Between BFI-2 and Mini-Markers. Study 2b replicated the patterns observed in Study 2a, with slightly lower correlation coefficients (see Table 7). For instance, for GPT-4o, the average correlations in Study 2a versus

Table 7*Correlation Between BFI-2 Input and LLM’s Mini-Marker Test Scores (Study 2b)*

Condition	Model	O	C	E	A	N	Avg
Human	Human	.75	.84	.88	.80	.74	.80
Simple Binary	GPT-3.5	.41/.73	.46/.57	.60/.87	.10/.30 [†]	.46/.54	.41/.60
Simple Binary	GPT-4	.46/.89	.61/.78	.65/.97	.27/.52	.76/.89	.55/.81
Simple Binary	GPT-4o	.48/.90	.62/.77	.65/.97	.23/.50	.75/.90	.55/.81
Simple Binary	Llama	.46/.88	.62/.78	.66/.98	.18/.41	.73/.87	.53/.78
Simple Binary	Deepseek	.45/.80	.58/.77	.66/.97	.26/.45	.75/.89	.54/.77
Elaborated Binary	GPT-3.5	.34/.76	.50/.67	.65/.95	.20/.31	.49/.59	.43/.66
Elaborated Binary	GPT-4	.45/.89	.62/.79	.66/.97	.28/.55	.75/.89	.55/.82
Elaborated Binary	GPT-4o	.47/.91	.61/.78	.65/.97	.28/.50	.77/.91	.56/.81
Elaborated Binary	Llama	.47/.91	.59/.77	.65/.98	.23/.47	.73/.87	.53/.80
Elaborated Binary	Deepseek	.50/.85	.61/.77	.66/.97	.23/.51	.75/.89	.55/.80
BFI-2-Likert	GPT-3.5	.51	.85	.86	.64	.71	.71
BFI-2-Likert	GPT-4	.68	.88	.92	.90	.81	.84
BFI-2-Likert	GPT-4o	.87	.92	.93	.88	.91	.90
BFI-2-Likert	Llama	.92	.90	.93	.86	.86	.89
BFI-2-Likert	Deepseek	.82	.92	.94	.87	.85	.88
BFI-2-Expanded	GPT-3.5	.83	.71	.87	.69	.72	.77
BFI-2-Expanded	GPT-4	.84	.84	.91	.82	.83	.85
BFI-2-Expanded	GPT-4o	.87	.92	.92	.90	.90	.90
BFI-2-Expanded	Llama	.88	.92	.92	.87	.88	.89
BFI-2-Expanded	Deepseek	.86	.92	.90	.84	.90	.88

Note. The sample size for Human and AI-Agents’ conditions are different (Human: $N=438$, AI-Agents: $N=200$). For Simple- and Elaborated-Binary conditions, there are two sets of correlations: (1) correlations between binary Big Five input scores and Mini-Markers output scores (right-hand values), and (2) correlations between original BFI-2 Likert scores and Mini-Markers output scores (left-hand values). For BFI-2-Likert and -Expanded conditions, correlations were between BFI-2 input and Mini-Markers output. All correlation coefficients were significantly different from zero at $\alpha = .05$, except where the dagger ([†]) indicates. Light-gray shading marks values *not* statistically different from Human reference after Bonferroni correction across domains ($\alpha = .05$; two-sided Fisher z test on Fisher-transformed r).

Study 2b were .55 vs. .59, .56 vs. .58, .90 vs. .91, and .90 vs. .90 for the Simple-Binary, Elaborated-Binary, BFI-2-Likert, and BFI-2-Expanded conditions, respectively. Similar patterns hold at the domain level. Consistent with Study 2a, correlations between Mini-Markers responses and the dichotomized binary BFI-2 input scores were uniformly higher than their corresponding correlations with the original BFI-2 scores. Additionally, the two more complex prompt conditions (BFI-2-Likert and BFI-2-Expanded), which performed at near-identical levels, consistently yielded higher correlations than the two baseline conditions (Simple-Binary and Elaborated-Binary).

In terms of human-AI alignment, Study 2b shows patterns very similar to those in Study 2a. Both BFI-2-Likert and BFI-2-Expanded showed good alignment, with GPT-3.5 and GPT-4 being the best-performing LLMs. For the Simple- and Elaborated-Binary conditions, the correlations between the original BFI-2 and Mini-Markers scores show very poor alignment with humans, except for Neuroticism; meanwhile, correlations between the binary Big Five input and Mini-Markers output scores showed good alignment for Openness and Extraversion but were inflated relative to human correlations for other domains.

Factor Analysis. Once again, the Study 2b factor analyses showed the same pattern of results as Study 2a⁵. Consistent with Study 2a, with the exception of GPT-3.5, LLMs exhibited unrealistically good fit and reliability under the Simple- and Elaborated-Binary prompt conditions, and poorer fit and lower reliability under the BFI-2-Likert and BFI-2-Expanded conditions. For example, GPT-4o obtained SRMR values of .02 (Simple-Binary), .02 (Elaborated-Binary), .11 (BFI-2-Likert), and .12 (BFI-2-Expanded), compared to a human reference of .09; and reliability coefficients of .99, .98, .89, and .84, compared to .83 for human reference. As documented in Study 2a, the binary prompt conditions yielded spuriously strong fit because newer LLMs closely track the low/high signals embedded in the prompts and consequently produce highly similar (often near-identical) responses across items within a domain, inflating inter-item correlations and, in turn, model-fit indices and reliability far beyond human

⁵ Full results from factor analyses are provided in the Supplementary Materials

levels. By contrast, the Likert and Expanded conditions preserved graded, continuous variation and subtler covariation among items, leading to loadings, fit indices, and reliability that more closely resemble human data (see Tables 5 and 8).

Overall, as shown by the results of item-level analyses, input-output correlational analyses and factor analyses, the strong parallelism between Study 2a and Study 2b suggests that creating AI-Agents using sample statistics provides a feasible alternative or precursor to collecting new human data. This method offers researchers a practical tool for exploring personality dynamics and interactions in a controlled, scalable environment, while maintaining psychometric validity comparable to traditional human-subject research.

Study 3: Further Validating AI-Agents with Real-Life Moral and Risk-Taking Vignettes

In Study 3, we aim to further validate AI-Agents with real-like moral and risk-taking vignettes and examine their alignment with human behaviors. This investigation seeks to elucidate both the potential and limitations of employing AI-Agents in behavioral science studies.

Methods

Measurement

In this study, we crafted ten vignettes: five risk-taking and five moral dilemma vignettes. The five risk-taking vignettes were designed to test an individual's risk-taking tendency versus risk-avoidance tendency. The specific vignettes included: 1) embarking on an entrepreneurial venture, 2) making significant investments, 3) confessing romantic feelings to a close friend, 4) participating in extreme sports, and 5) opting to study overseas (See Appendix II). For the risk-taking vignettes, higher scores indicate higher risk-seeking and lower risk-avoidance.

The five moral dilemma vignettes were designed to assess individuals' empathetic tendency versus rule adherence tendency. Given that the majority of currently available LLMs have been safety-aligned and instruction fine-tuned (Biedma et al., 2024), which typically skews them away from engaging in severe moral judgments

Table 8
Fit Measures for CFA Models in Study 2a

Fit Measure	Human						Simple-Binary						Elaborated-Binary						BFI-2-Likert						BFI-2-Expanded																			
	Ref			3.5	4	40	Llama	DS	3.5			4	40	Llama	DS	3.5			4	40	Llama	DS	3.5			4	40	Llama	DS															
	Chi-Squared	238.59	587.55	323.25	833.66	741.86	163.11	562.79	211.50	664.82	728.75	144.12	146.38	395.06	668.01	558.96	596.41	754.33	471.32	906.42	433.74	444.58	Chi-Squared	238.59	587.55	323.25	833.66	741.86	163.11	562.79	211.50	664.82	728.75	144.12	146.38	395.06	668.01	558.96	596.41	754.33	471.32	906.42	433.74	444.58
Extraversion	CFI	.87	.93	.97	.93	.95	.88	.93	.98	.94	.95	.89	.87	.86	.87	.85	.84	.87	.82	.92	.90	.90	CFI	.87	.93	.97	.93	.95	.88	.93	.98	.94	.95	.89	.87	.86	.87	.85	.84	.87	.82	.92	.90	.90
	RMSEA	.16	.26	.19	.30	.29	.43	.25	.15	.27	.28	.40	.41	.21	.27	.25	.26	.29	.23	.32	.22	.22	RMSEA	.16	.26	.19	.30	.29	.43	.25	.15	.27	.28	.40	.41	.21	.27	.25	.26	.29	.23	.32	.22	.22
	SRMR	.07	.03	.01	.01	.00	.01	.03	.00	.01	.00	.01	.07	.07	.10	.07	.05	.03	.04				SRMR	.07	.03	.01	.01	.00	.01	.03	.00	.01	.00	.01	.07	.07	.10	.09	.05	.03	.04			
Agreeableness	Chi-Squared	225.69	562.12	333.68	975.38	3244.27	2822.90	572.83	381.25	1082.12	4006.49	2627.98	1236.72	578.75	873.35	1232.21	754.11	895.40	682.44	695.87	1209.57	1211.47	Chi-Squared	225.69	562.12	333.68	975.38	3244.27	2822.90	572.83	381.25	1082.12	4006.49	2627.98	1236.72	578.75	873.35	1232.21	754.11	895.40	682.44	695.87	1209.57	1211.47
	CFI	.84	.80	.95	.84	.60	.70	.79	.94	.83	.53	.71	.63	.81	.78	.74	.79	.73	.76	.82	.72	.69	CFI	.84	.80	.95	.84	.60	.70	.79	.94	.83	.53	.71	.63	.81	.78	.74	.79	.73	.76	.82	.72	.69
	RMSEA	.15	.25	.19	.33	.61	.57	.25	.20	.35	.68	.55	.37	.25	.31	.29	.32	.28	.37	.31	.37	.37	RMSEA	.15	.25	.19	.33	.61	.57	.25	.20	.35	.68	.55	.37	.25	.31	.29	.32	.28	.37	.31	.37	.37
	SRMR	.08	.17	.03	.06	.14	.12	.19	.03	.06	.13	.14	.29	.11	.08	.10	.09	.18	.09	.07	.08	.09	SRMR	.08	.17	.03	.06	.14	.12	.19	.03	.06	.13	.14	.29	.11	.08	.10	.09	.18	.09	.07	.08	.09
Conscientiousness	Chi-Squared	317.88	1147.92	202.22	357.77	1409.29	1356.58	1252.59	154.42	453.51	1388.33	877.57	767.39	631.95	626.56	723.88	759.83	916.01	668.71	536.44	915.88	901.91	Chi-Squared	317.88	1147.92	202.22	357.77	1409.29	1356.58	1252.59	154.42	453.51	1388.33	877.57	767.39	631.95	626.56	723.88	759.83	916.01	668.71	536.44	915.88	901.91
	CFI	.79	.98	.96	.85	.85	.85	.67	.98	.95	.85	.90	.62	.69	.80	.79	.70	.71	.72	.84	.79	.74	CFI	.79	.98	.96	.85	.85	.85	.67	.98	.95	.85	.90	.62	.69	.80	.79	.70	.71	.72	.84	.79	.74
	RMSEA	.18	.36	.14	.20	.40	.39	.38	.12	.22	.40	.31	.29	.26	.26	.28	.29	.32	.27	.32	.32	.32	RMSEA	.18	.36	.14	.20	.40	.39	.38	.12	.22	.40	.31	.29	.26	.26	.28	.29	.32	.27	.32	.32	.32
	SRMR	.09	.14	.01	.01	.03	.03	.17	.01	.02	.03	.02	.20	.14	.08	.11	.10	.11	.07	.08	.08	.08	SRMR	.09	.14	.01	.01	.03	.03	.17	.01	.02	.03	.02	.20	.14	.08	.11	.10	.11	.07	.08	.08	.08
Negative Emotionality	Chi-Squared	329.78	1484.58	1233.61	477.87	1628.75	1384.60	1494.83	1273.36	656.71	1818.51	1083.63	775.23	798.21	100.01	1176.54	1062.65	1646.45	966.10	1122.56	1656.90	1327.83	Chi-Squared	329.78	1484.58	1233.61	477.87	1628.75	1384.60	1494.83	1273.36	656.71	1818.51	1083.63	775.23	798.21	100.01	1176.54	1062.65	1646.45	966.10	1122.56	1656.90	1327.83
	CFI	.78	.36	.83	.96	.78	.86	.38	.83	.94	.76	.88	.75	.61	.70	.70	.72	.42	.58	.72	.59	.67	CFI	.78	.36	.83	.96	.78	.86	.38	.83	.94	.76	.88	.75	.61	.70	.70	.72	.42	.58	.72	.59	.67
	RMSEA	.19	.41	.38	.23	.43	.40	.41	.38	.27	.45	.35	.29	.30	.33	.36	.34	.43	.33	.35	.43	.39	RMSEA	.19	.41	.38	.23	.43	.40	.41	.38	.27	.45	.35	.29	.30	.33	.36	.34	.43	.33	.35	.43	.39
	SRMR	.10	.28	.14	.01	.12	.03	.13	.01	.12	.04	.04	.20	.17	.12	.11	.19	.24	.13	.21	.21	.21	SRMR	.10	.28	.14	.01	.12	.03	.13	.01	.12	.04	.04	.20	.17	.12	.11	.19	.24	.13	.21	.21	.21
Open-Mindedness	Chi-Squared	418.21	449.04	365.43	791.24	1513.55	1754.92	443.68	344.24	752.06	1596.25	120.19	846.06	152.88	1659.83	1914.50	2266.87	1551.01	2458.43	2189.52	2053.83	2841.07	Chi-Squared	418.21	449.04	365.43	791.24	1513.55	1754.92	443.68	344.24	752.06	1596.25	120.19	846.06	152.88	1659.83	1914.50	2266.87	1551.01	2458.43	2189.52	2053.83	2841.07
	CFI	.71	.83	.93	.89	.75	.77	.85	.94	.90	.74	.83	.66	.58	.59	.60	.53	.50	.51	.54	.54	.54	CFI	.71	.83	.93	.89	.75	.77	.85	.94	.90	.74	.83	.66	.58	.59	.60	.53	.50	.51	.54	.54	.54
	RMSEA	.21	.22	.20	.30	.41	.45	.22	.19	.42	.37	.31	.41	.43	.42	.47	.51	.42	.53	.50	.48	.57	RMSEA	.21	.22	.20	.30	.41	.45	.22	.19	.42	.37	.31	.41	.43	.42	.47	.51	.42	.53	.50	.48	.57
	SRMR	.12	.12	.03	.09	.07	.11	.03	.09	.05	.13	.28	.14	.13	.18	.24	.22	.20	.26			SRMR	.12	.12	.03	.09	.07	.11	.03	.09	.05	.13	.28	.14	.13	.18	.24	.22	.20	.26				

Note. Column blocks list models as: 3.5 = GPT-3.5 Turbo (0125); 4 = GPT-4; 40 = GPT-40; Llama = Llama-3.30-70B-Instruct; DS = DeepSeek-V3. “Human” is the reference fit; “Ref” is its summary column. Averages are computed across the five subscales within each condition model. Values rounded to 2 decimals (leading zero omitted). The asterisk beside the chi-square for Open-Mindedness \times GPT-4 \times BFI-2-Likert indicates that the model fit yielded a warning (Heywood case). Light gray = acceptable fit ($CFI \geq .90$, $RMSEA \leq .08$, $SRMR \leq .08$); dark gray = very good fit ($CFI \geq .95$, $RMSEA \leq .05$, $SRMR \leq .05$).

(e.g., the Trolley Problem), we introduced a series of everyday moral dilemmas. These dilemmas necessitated choosing between upholding a standard and prioritizing empathy. Examples include decisions on 1) reporting a friend for cheating on a quiz, 2) addressing a colleague's misappropriation of office supplies, 3) underage drinking, 4) disclosing confidential information that could save lives, and 5) providing candid feedback on subpar performance (see Appendix III). For the moral dilemmas vignettes, higher scores indicate higher rule adherence tendency and lower empathetic tendency.

Procedure

For human data, we conducted a survey study at a Canadian University. Participants were asked to complete the BFI-2 Expanded-format scale and respond to the ten vignettes. After applying exclusion criteria based on English proficiency, consent for data deposit, and survey completion, we retained a sample of 276 participants. The mean age of participants was 19.65 years ($SD = 3.88$). Regarding gender identity, the majority of participants identified as female (80.4%), with 15.6% identifying as male and 4% identified as Other. The ethnic composition of the sample was diverse: 26.8% as European/Caucasian, 25.0% as South Asian, 7.3% as African, 6.9% as East Asian, 2.5% as Latino and Hispanic and 28.3% identified as Other.

To create the AI-Agents, same as Study 2, first, we used the four prompting strategies (Simple-Binary, Elaborated-Binary, BFI-2-Likert and BFI-2-Expanded, see Table 1) and five LLMs (GPT-3.5, GPT-4, GPT-4o, Llama, and DeepSeek). For each prompt \times LLM condition, we created 276 AI-Agents, each corresponding to a human participant's personality. We then prompted the AI-Agents to read five risk-taking vignettes and five moral dilemmas and to indicate their decisions on a 1-10 scale (see Appendices C and D for prompts).

Data Analyses

Similar to Study 2, in Study 3, we conducted analyses comparing human and AI-Agents responses on the ten vignettes. We compared the means and distributions of their responses using paired-wise t -test and KS tests, respectively. In addition, we conducted regression analyses in which the original Big Five trait scores assigned to the

AI-Agents were used to predict their responses to the ten vignettes. All significance tests employed Bonferroni corrections to control the family-wise Type I error rate.

Results and Discussion

Comparing Human and AI-Agents Responses on the Risk and Moral

Vignettes. Tables 9 and 10 show the mean and distribution differences between AI-Agents and human responses for the ten vignettes. Figure 4 shows graphic representations of these differences in selected conditions.

The most noticeable pattern of results is that the newer LLMs are clearly fine-tuned to provide morally acceptable responses in the moral dilemma vignettes. As shown in Table 9 and Figure 4, for dilemmas involving confidential information, underage drinking, and exam cheating, AI-Agents created with newer LLMs consistently scored substantially higher across prompt conditions, with differences ranging from 1.02 to 3.88. As a result, the AI-Agents' score distributions are heavily skewed and significantly different from those of humans (see Table 10 and Figure 4). In contrast, AI-Agents created by the older LLM (GPT-3.5) often scored lower on moral vignettes than human participants. This clear divergence suggests that the newer LLMs have been deliberately fine-tuned for safety, aligning their responses more closely with socially desirable moral norms. This pattern of results is consistent with our expectation that the newer LLMs have been deliberately fine-tuned for safety reasons.

For risk-taking vignettes, the mean and distributional differences between AI-Agents and human responses were considerably lower, especially in the BFI-2-Likert and -Expanded conditions (see Tables 9 and 10 and Figure 4). In fact, for the investment and confession vignettes, AI-Agents produced mean scores that were comparable to those of humans (as indicated by the shaded cells in Table 9). However, even in these cases, the responses of AI-Agents, especially those generated by the newer LLMs, displayed substantially less variation than human responses (see Figure 4). In fact, the investment and confession risk-taking vignette had similar mean scores compared to humans (as shown by the shaded cells in Table 9), although the AI-Agents' responses still showed considerably less variation than human responses for the newer

LLMs (see Figure 4).

Comparing across the prompt conditions, for the moral dilemma vignettes, all prompt conditions yielded responses that were heavily skewed to the right (see Figure 4). On the other hand, for risk-taking vignettes, the AI-Agents in the BFI-2-Likert and -Expanded conditions generated more variable responses, whereas those in the Simple- and Elaborated-Binary conditions continued to display skewed distributions.

Table 9

Mean Differences between AI-Agents and Human Responses on the Risk and Moral Vignettes

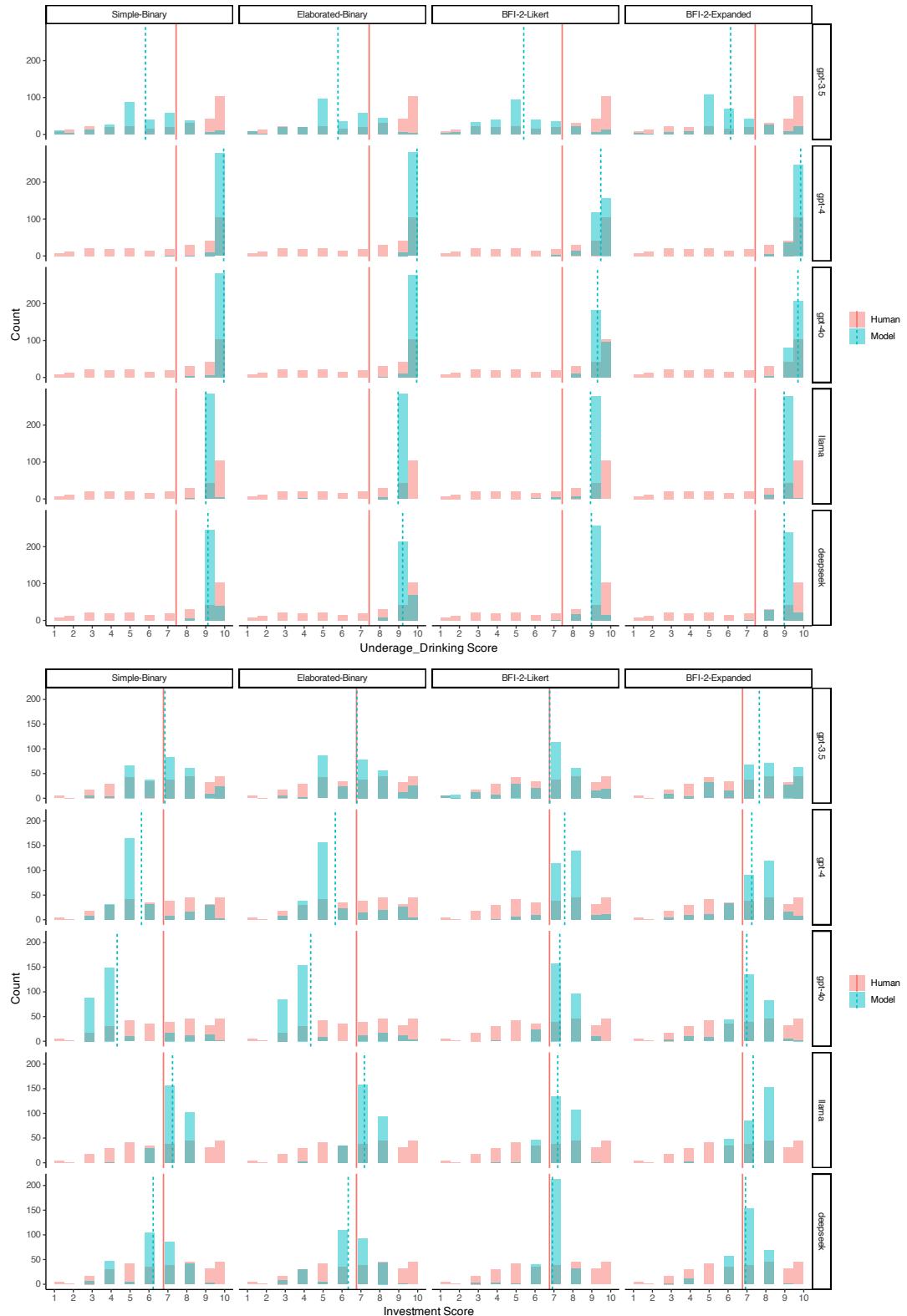
Condition	LLM	Conf Info	Under Age	Exam Cheat	Honest FB	Work Theft	Invest	Ext Sports	Entre Vent	Confess	Study Abroad
Simple-Binary	GPT-3.5	-0.74	-1.62	1.02	0.49	0.76	0.08	-2.68	-2.86	0.23	-2.27
Simple-Binary	GPT-4	3.88	2.51	4.98	1.12	2.09	-1.16	1.50	-0.99	-1.70	-1.10
Simple-Binary	GPT-4o	2.15	2.52	3.53	0.78	3.15	-2.45	0.33	-1.69	-2.26	-1.39
Simple-Binary	Llama	2.16	1.57	3.84	1.03	2.93	0.48	1.18	-0.12	-1.37	-0.69
Simple-Binary	DeepSeek	1.93	1.68	2.93	0.34	2.18	-0.54	1.09	-1.28	-1.60	-0.83
Elaborated-Binary	GPT-3.5	-0.47	-1.65	1.04	0.44	0.81	0.02	-2.60	-2.98	0.34	-2.43
Elaborated-Binary	GPT-4	3.84	2.53	4.84	1.03	2.02	-1.12	1.49	-0.92	-1.64	-1.11
Elaborated-Binary	GPT-4o	2.09	2.51	3.57	0.76	3.18	-2.42	0.22	-1.74	-2.28	-1.40
Elaborated-Binary	Llama	2.12	1.53	3.84	1.01	2.89	0.42	1.20	-0.10	-1.33	-0.71
Elaborated-Binary	DeepSeek	1.96	1.77	3.00	0.47	2.19	-0.44	1.24	-1.17	-1.60	-0.98
BFI-2-Likert	GPT-3.5	-1.49	-2.03	-0.64	-0.90	0.53	0.01	-2.56	-3.24	-0.39	-2.13
BFI-2-Likert	GPT-4	2.93	2.04	4.21	1.23	3.13	0.80	2.37	1.83	0.11	1.63
BFI-2-Likert	GPT-4o	1.44	1.87	3.06	-0.97	2.74	0.54	2.22	1.73	0.86	1.41
BFI-2-Likert	Llama	2.09	1.50	3.86	0.85	3.21	0.43	0.94	0.66	-0.03	1.09
BFI-2-Likert	DeepSeek	2.10	1.54	2.89	0.16	2.13	0.15	1.53	-0.03	-0.68	0.66
BFI-2-Expanded	GPT-3.5	-2.52	-1.30	-0.80	-0.72	0.79	0.89	-0.77	-0.91	1.28	-0.37
BFI-2-Expanded	GPT-4	3.52	2.40	4.31	0.83	2.20	0.49	2.60	1.23	-0.10	0.80
BFI-2-Expanded	GPT-4o	1.81	2.26	3.12	-0.28	2.88	0.22	2.08	0.81	-0.06	0.17
BFI-2-Expanded	Llama	2.08	1.53	3.48	0.50	3.27	0.56	0.82	0.21	-0.44	1.51
BFI-2-Expanded	DeepSeek	2.03	1.54	2.62	0.12	2.12	0.16	1.51	-0.23	-0.68	0.55

Note: Values are AI – Human mean differences. Unshaded cells indicate significant paired-sample *t*-tests (Bonferroni-corrected); gray cells indicate non-significant results. Conf Info=Vignette about disclosing confidential information; Under Age=Vignette about underage drinking; Exam Cheat=Vignette about exam cheating; Honest FB=Vignette about providing candid feedback on subpar performance; Worth Theft=Vignette about workplace theft; Invest=Vignette about making investment; Ext Sports=Vignette about participating in extreme sports; Entre Vent=Vignette about embarking on an entrepreneurial venture; Confess=Vignette about confessing to a close friend; Study Abroad=Vignette about studying abroad.

Predicting Risk and Moral Vignettes Scores Based on Big Five. Table

11 of standardized regression coefficients shows how each Big Five domain predicts moral- and risk-taking decisions across response formats and models. Among human participants, for the risk-taking vignettes, individuals higher in Openness and Extraversion were significantly more likely to endorse risky actions ($O: \beta = .14^*$, $E: \beta = .23^*$), whereas higher Neuroticism predicted a greater avoidance of risk ($\beta = -.15^*$). Conscientiousness and Agreeableness show no significant associations with risk. For the moral dilemma vignettes, Conscientiousness positively predicted prosocial choices ($\beta =$

Figure 4
Selected Scores on Vignettes For Human and AI-Agents



Note. The red line indicates the mean of the human responses. The dotted blue lines indicate the means of AI-Agents across conditions.

Table 10

Kolmogorov-Smirnov Test Statistics Comparing AI-Agents and Human Responses on the Risk and Moral Vignettes

Condition	LLM	Conf Info	Under Age	Exam Cheat	Honest FB	Work Theft	Invest	Ext Sports	Entre Vent	Confess	Study Abroad
Simple-Binary	GPT-3.5	0.18	0.45	0.26	0.20	0.43	0.16	0.58	0.67	0.27	0.49
Simple-Binary	GPT-4	0.78	0.60	0.86	0.28	0.49	0.38	0.35	0.36	0.41	0.29
Simple-Binary	GPT-4o	0.54	0.61	0.79	0.40	0.74	0.63	0.29	0.48	0.44	0.39
Simple-Binary	Llama	0.70	0.49	0.86	0.49	0.80	0.34	0.33	0.24	0.48	0.26
Simple-Binary	DeepSeek	0.60	0.48	0.79	0.43	0.70	0.27	0.33	0.38	0.38	0.22
Elaborated-Binary	GPT-3.5	0.14	0.47	0.29	0.19	0.41	0.16	0.58	0.69	0.29	0.51
Elaborated-Binary	GPT-4	0.76	0.61	0.84	0.29	0.50	0.37	0.35	0.37	0.40	0.30
Elaborated-Binary	GPT-4o	0.53	0.60	0.80	0.40	0.72	0.64	0.30	0.50	0.46	0.40
Elaborated-Binary	Llama	0.69	0.48	0.87	0.49	0.80	0.32	0.35	0.24	0.45	0.27
Elaborated-Binary	DeepSeek	0.59	0.47	0.76	0.42	0.69	0.27	0.36	0.37	0.40	0.23
BFI-2-Likert	GPT-3.5	0.28	0.48	0.22	0.28	0.32	0.17	0.49	0.73	0.26	0.48
BFI-2-Likert	GPT-4	0.60	0.44	0.76	0.25	0.57	0.39	0.50	0.50	0.32	0.47
BFI-2-Likert	GPT-4o	0.44	0.46	0.72	0.41	0.68	0.36	0.52	0.44	0.32	0.47
BFI-2-Likert	Llama	0.66	0.46	0.87	0.43	0.85	0.32	0.27	0.30	0.33	0.47
BFI-2-Likert	DeepSeek	0.62	0.43	0.74	0.37	0.69	0.31	0.48	0.30	0.34	0.40
BFI-2-Expanded	GPT-3.5	0.50	0.42	0.22	0.24	0.33	0.24	0.20	0.40	0.31	0.25
BFI-2-Expanded	GPT-4	0.71	0.50	0.83	0.24	0.52	0.25	0.58	0.24	0.26	0.33
BFI-2-Expanded	GPT-4o	0.46	0.48	0.72	0.31	0.71	0.26	0.47	0.24	0.21	0.30
BFI-2-Expanded	Llama	0.67	0.46	0.86	0.40	0.74	0.32	0.24	0.22	0.28	0.47
BFI-2-Expanded	DeepSeek	0.54	0.40	0.74	0.26	0.62	0.29	0.41	0.24	0.33	0.33

Note: Values are KS D statistics. Unshaded cells indicate significant KS tests (Bonferroni-corrected); gray cells indicate non-significant results. Conf Info=Vignette about disclosing confidential information; Under Age=Vignette about underage drinking; Exam Cheat=Vignette about exam cheating; Honest FB=Vignette about providing candid feedback on subpar performance; Worth Theft=Vignette about workplace theft; Invest=Vignette about making investment; Ext Sports=Vignette about participating in extreme sports; Entre Vent=Vignette about embarking on an entrepreneurial venture; Confess=Vignette about confessing to a close friend; Study Abroad=Vignette about studying abroad.

.21***), while Neuroticism negatively predicted them ($\beta = -.15^*$); the remaining domains were not significant. These human patterns provided a reference for interpreting AI-Agent behavior.

Risk-Taking Vignettes Scores. For risk-taking, AI-Agents created with Simple- and Elaborated-Binary and BFI-2-Expanded were all good at recovering human prediction patterns (i.e., direction and significance). Across simple- and elaborated-binary and the BFI-2-Expanded prompt condition, AI-Agents displayed a common pattern that mirrored human risk-taking behavior. In all three conditions, coefficients for Openness and Extraversion were large and positive, while Neuroticism was strongly negative. These effects were much larger than those observed in humans and remained significant across most models. It suggests that AI-Agents consistently linked curiosity and sociability with greater risk appetite and associate emotional instability with risk aversion.

Meanwhile, the BFI-2-Expanded prompt condition differed from two baseline

conditions: the expanded-prompt elicited larger Extraversion coefficients ($\beta = 0.65\text{--}0.81$) and Neuroticism coefficients ($\beta = -.29\text{--}-.41$) but Openness coefficients that were slightly smaller or comparable to those in the baseline (expanded $\beta = .19\text{--}.35$ vs. baseline $\beta = .26\text{--}.41$), and it attenuated inconsistent and spurious effects of other domains. In contrast, the baseline binary prompts sometimes produced small but significant positive Agreeableness and Conscientiousness coefficients (e.g., GPT-4 A=0.20*** in simple-binary; GPT-4 C=0.15*). Thus, both prompting approaches captured the human-like pattern of risk-taking, but the expanded-prompt condition provided a cleaner representation; the main exception was that several GPT models continue to show modest positive effects for Agreeableness in the baseline conditions.

In contrast, under the BFI-2-Likert condition, coefficient patterns diverged from the human reference. Only isolated recoveries appeared: Openness was positive only for GPT-3.5 ($\beta = .13^*$), Extraversion only for Llama ($\beta = .31^*$), and the negative Neuroticism effect for GPT-3.5 ($\beta = -.13^*$), GPT-4o ($\beta = -.13^*$), and Llama ($\beta = -.20^{***}$). Signs and magnitudes for the remaining domains frequently flipped or attenuated, and this condition did not reliably reproduce the human non-significant associations for Conscientiousness and Agreeableness. Overall, unlike the BFI-2-Expanded and binary prompt conditions, the Likert condition failed to recover most of the associations between personality and risk-taking behaviors.

Moral Dilemma Scores. Patterns for moral dilemma are similar to risk-taking patterns. In binary and expanded prompt conditions, AI-Agents' decisions largely followed the human pattern: prosocial choices increased with higher Conscientiousness and decreased with higher Neuroticism, whereas the other domains showed little effect. Expanded prompt conditions outperformed the two binary prompt conditions, as they more consistently reproduced the significant positive coefficients for Conscientiousness and significant positive coefficients for Neuroticism (except for GPT-3.5 and GPT-4) while keeping the non-significant coefficients for the other three domains. These results show that the expanded-prompt condition produced clearer and stronger effects while attenuating spurious effects and yielding a closer match to human

data.

In contrast, the BFI-2-Likert prompt condition shows much worse alignment with the human reference. It only fully recovered the non-significant coefficients for Openesses, and partially recovers the significant positive coefficients for Conscientiousness (GPT-3.5: $\beta=.18^{**}$, GPT-4: $\beta=.23^{***}$, and Deepseek: $\beta=.17^{**}$), while the rest of domains show inconsistent patterns that fail to match the human reference.

Model comparisons across Risk-taking and Moral Dilemma vignettes.

Across both risk-taking and moral dilemma vignettes, model performance varied dramatically between the BFI-2-Likert and BFI-2-Expanded conditions. In the BFI-2-Expanded prompt condition, GPT-4o demonstrated the best alignment with human reference, followed by GPT-4 and Llama, showing the most robust reproduction of human patterns with consistently large positive coefficients for Openness (risk: $\beta=.32^{***}$) and Extraversion (risk: $\beta=.81^{***}$; moral: non-significant as expected), and strong negative Neuroticism effects across both vignette types.

The BFI-2-Likert condition revealed striking model differences and generally poor performance. Most concerningly, GPT-4 and GPT-4o showed patterns directly opposite to human reference in this format, producing negative coefficients where humans showed positive ones (e.g., GPT-4o moral Extraversion: $\beta=-.21^{***}$ vs. human non-significant). GPT-3.5 and Llama maintained better directional consistency in the Likert format, with GPT-3.5 partially recovering human risk patterns (Openness: $\beta=.13^*$) and Llama showing robust Extraversion effects (risk: $\beta=.31^{***}$). DeepSeek showed intermediate performance, with some preserved directional relationships but inconsistent magnitudes.

Overall, newer, more sophisticated models (GPT-4, GPT-4o) showed better format sensitivity, performing very well in the BFI-2-Expanded condition but poorly in the BFI-2-Likert condition. This suggests that advanced models may be more susceptible to prompt-specific linguistic cues. Meanwhile, in BFI-2-Expanded condition, AI-Agent consistently showed amplified true effects while minimizing spurious

associations across all models, indicating its superior validity for simulating personality-behavior associations.

One plausible account of the BFI-2-Likert format's weaker behavioral prediction, even relative to the binary baselines, is its indirectness. In the Likert condition, trait information is conveyed numerically, so the model must first translate numbers into semantic descriptions before mapping those meanings onto behavioral choices, introducing an extra decoding step. By contrast, both BFI-2-Expanded and the binary baselines present trait cues directly in natural language, which the models appear to process more fluently in this context. This mechanism is consistent with the observed pattern that Likert underperforms while language-native inputs yield cleaner, more stable behavior mappings.

Finally, across models and vignettes, BFI-2-Expanded generally attained the strongest behavioral prediction. We interpret this advantage as the result of Expanded prompts providing richer, sentence-level semantics that minimize representational indirection and better leverage language-native processing. The binary baselines, while more intuitive than Likert, sacrifice granularity; the Likert format, while granular, imposes a numeric-to-semantic translation burden. Taken together, these results suggest that, in this setting, expressing trait information as full sentences (Expanded) offers the most effective and human-like linkage between measured traits and predicted behaviors.

General Discussion

The present research introduces a novel methodology for assigning quantifiable, controllable, and psychometrically validated personalities to Agents using the Big Five personality framework. Through a series of three studies, we demonstrated both the potential and the limitations of this approach for social science research.

Study 1 demonstrated semantic similarities between different Big Five personality measures within the embedding space of LLMs. This finding suggests that LLMs are capable of interpreting and representing personality-related concepts, providing the basis for subsequent studies aimed at validating AI-Agents with assigned Big Five personalities.

Table 11*Standardized Coefficients For Predicting the Risk-Taking and Moral Dilemma Vignettes*

Vignette	Format	Model	Predictor				
			O	C	E	A	N
Risk	Simple-Binary	human	0.14*	-0.04	0.24***	-0.03	-0.19 **
		GPT-3.5	0.34***	0.06	0.51***	0.14*	-0.15 *
		GPT-4	0.32***	0.15*	0.49***	0.20***	-0.36 ***
		GPT-4o	0.41***	0.09	0.52***	0.17**	-0.21 ***
		Llama	0.27***	0.03	0.46***	0.13*	-0.30 ***
	Elaborated-Binary	Deepseek	0.28***	0.11	0.48***	0.14*	-0.32 ***
		GPT-3.5	0.31***	0.08	0.52***	0.19**	-0.16 **
		GPT-4	0.32***	0.15*	0.52***	0.22***	-0.30 ***
		GPT-4o	0.39***	0.07	0.50***	0.17**	-0.21 ***
		Llama	0.31***	0.06	0.50***	0.18**	-0.31 ***
	BFI-2-Likert	Deepseek	0.26***	0.16**	0.42***	0.18**	-0.34 ***
		GPT-3.5	0.13*	0.23***	0.11	0.06	-0.13 *
		GPT-4	-0.05	-0.12 *	0.08	-0.13 *	-0.01
		GPT-4o	-0.07	-0.08	0.10	-0.16 **	-0.13 *
		Llama	0.08	0.17**	0.31***	-0.08	-0.20 ***
	BFI-2-Expanded	Deepseek	0.08	0.07	0.02	-0.19 ***	-0.05
		GPT-3.5	0.19**	0.02	0.65***	0.13*	-0.29 ***
		GPT-4	0.23***	0.07	0.76***	0.17**	-0.35 ***
		GPT-4o	0.32***	0.14*	0.81***	0.19**	-0.41 ***
		Llama	0.35***	0.03	0.73***	-0.08	-0.24 ***
		Deepseek	0.25***	0.09	0.69***	0.02	-0.36 ***
Moral	Simple-Binary	human	-0.03	0.21***	0.01	0.05	-0.15 *
		GPT-3.5	-0.03	0.10	0.15*	0.10	-0.11
		GPT-4	-0.02	0.24***	0.15*	0.07	-0.31 ***
		GPT-4o	-0.12 *	0.35***	-0.06	0.03	-0.25 ***
		Llama	-0.14 *	0.11	-0.03	0.08	-0.23 ***
	Elaborated-Binary	Deepseek	-0.02	0.13*	-0.06	0.03	-0.04
		GPT-3.5	0.02	0.04	0.17**	0.09	-0.14 *
		GPT-4	-0.06	0.11	-0.02	0.06	-0.17 **
		GPT-4o	-0.15 *	0.31***	-0.04	0.04	-0.32 ***
		Llama	-0.16 **	0.18**	0.07	0.08	-0.21 ***
	BFI-2-Likert	Deepseek	-0.09	0.27***	0.05	0.02	-0.25 ***
		GPT-3.5	0.04	0.18**	0.17**	0.00	-0.01
		GPT-4	0.07	0.23***	0.14*	0.16**	-0.08
		GPT-4o	-0.11	-0.13 *	-0.21 ***	-0.31 ***	0.22***
		Llama	0.00	-0.02	-0.07	-0.07	0.00
	BFI-2-Expanded	Deepseek	0.03	0.17**	-0.08	0.17**	0.04
		GPT-3.5	-0.06	0.16**	0.02	-0.20 ***	-0.08
		GPT-4	0.01	0.24***	-0.07	-0.04	-0.10
		GPT-4o	0.04	0.57***	-0.03	0.02	-0.31 ***
		Llama	-0.04	0.51***	0.06	0.12*	-0.28 ***
		Deepseek	-0.05	0.38***	0.04	0.07	-0.23 ***

Note. Entries are standardized coefficients (two decimals; leading zero omitted). * $p < .05$, ** $p < .01$, *** $p < .001$. O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism.

In Study 2, we designed a pipeline to create AI-Agents with personalities using prompts in the BFI-2 Likert and Expanded formats. We then validate these AI-Agents by examining their alignment with human participants' responses on a criterion measure, the Mini-Markers test, and compare them to AI-Agents created with binary adjective prompts (i.e., the Simple- and Elaborated-Binary conditions). Our results show that compared to AI-Agents created with binary adjective prompts or old LLM (i.e., GPT-3.5), AI-Agents created with the BFI-2-Likert and -Expanded prompts and with newer and more capable LLMs (i.e., GPT-4, GPT-4o, Llama, and DeepSeek) are much more aligned with human responses in terms of item means, item distributions, factor structure, reliability, and correlations between the BFI-2 input and the Mini-Markers outputs. However, this alignment is far from perfect. For example, although the average factor loading magnitudes were similar between AI-Agents (in BFI-2-Likert and -Expanded conditions with newer LLMs) and humans, the specific patterns of loading sizes (e.g., which items had the highest loadings) are quite different.

In Study 3, we further validated the AI-Agents with real-life decision-making vignettes describing risk-taking situations and moral dilemmas. For predicting decisions from personalities, AI-Agents created with the BFI-2-Expanded prompts achieved the strongest and most human-like behavior prediction, while BFI-2-Likert led to the worst performing AI-Agent. One possible explanation is that natural-language inputs (BFI-2-Expanded) reduce representational indirectness relative to numeric inputs (BFI-2-Likert), suggesting the Expanded prompt has the best potential in creating AI-Agents with realistic profiles. For the item-level response pattern, consistent Study 2, AI-Agents created with the BFI-2-Likert or -Expanded prompts and the newer LLMs outperformed those created with binary adjectives. Nevertheless, as expected, newer LLMs had clearly been fine-tuned for safety, resulting in AI-Agents that consistently produced morally inclined responses to the dilemma vignettes, regardless of their assigned personalities. Interestingly, even though these AI-Agents' responses were heavily skewed towards the high end of the moral scale (typically 8-10 on a 1-10 scale), correlations between Big Five traits and moral dilemma scores closely mirror those

observed in human participants. This pattern of results suggests that while fine-tuning shifted the overall response distribution towards more morally inclined responses, the underlying relationships between variables (e.g., between Big Five traits and moral dilemma responses) remain relatively unaffected.

Overall, our findings demonstrate both the potential and limitations of using AI-Agents as stand-ins for human participants in psychological research. On the one hand, alignment between AI-Agents and humans, particularly in correlations between input Big Five traits and output responses, suggests that AI-Agents hold strong potential as useful tools for preliminary investigations or pilot studies. On the other hand, the discrepancies we observed, such as skewed responses in the moral dilemma vignettes and divergences in the finer patterns of results, underscore that AI-Agents cannot fully substitute for human participants when drawing inferences in large-scale research projects yet.

Limitations and Future Directions

There are several limitations to our study. One limitation is that our method of simulating AI-Agents using only prompts designed from the BFI-2 falls short of capturing the intricate interplay of other important personality constructs, as well as individual backgrounds and societal complexity. We acknowledge that personality traits, while central to much of personality psychology, represent only one facet of what most people consider to be “personality.” Traits capture relatively stable patterns of thinking, feeling, and behaving, but they do not fully encompass other key components such as values, goals, roles, social identity, and emotional dynamics. Recent work has begun to investigate these other aspects in LLMs, such as how LLMs represent emotions (Li et al., 2023), human values (Yao et al., 2025), or identity (A. Wang et al., 2025). These studies highlight that personality in artificial agents can, in principle, extend far beyond trait-level description. Building on these studies, a more integrative approach to LLM personality simulation could combine trait-based control with frameworks for modeling values, motivational systems, and identity-related processes. Such an approach would allow researchers to capture not only how an agent behaves

across contexts, but also why it behaves that way, and how it might adapt in response to changing goals or social roles. Expanding in this direction could make AI-Agents a richer platform for studying complex aspects of human personality, bridging dispositional, motivational, and narrative analyses.

Another important limitation concerns the distinction between genuine personality expression and role-playing in AI-Agents. We assigned personality traits to AI-Agents and instructed them to enact these roles using validated psychometric tools. This design choice avoids claims that AI-Agents possess intrinsic or genuine personalities; instead, we leverage their capacity to simulate assigned traits with fidelity and experimental control. We acknowledge that this role-playing framework may limit generalizability. The AI-Agents' trait expression is context-dependent, shaped by the prompts and scenarios provided, and may not extend beyond those settings. Our primary focus is on the methodological utility of AI-Agents as controllable and transparent tools for simulating personality variation in structured research contexts. We do not make claims about consciousness, subjective experience, or "humanhood" in AI, which are questions that remain open for philosophical and interdisciplinary debate. Instead, we position our work as a practical contribution to personality and behavioral research, while acknowledging its limitations and the broader questions that lie beyond its scope.

Finally, due to the large number of conditions in our existing studies, we did not examine LLM parameters such as temperature and top-p; these were kept at their default values across our studies. Theoretically, these parameters could be optimized to generate responses that more closely align with human data. We are currently addressing this in a follow-up study on AI-Agents, using machine learning techniques to optimize these settings and improve response quality.

Conclusion

In conclusion, this research presents a novel and promising approach to creating AI-Agents with psychometrically valid personality traits. Through a series of studies, we showed that AI-Agents created with newer LLMs and BFI-2 prompts have the

potential to be used as stand-ins for primary investigations and pilot studies, especially for examining the relationships between personality-related variables. However, AI-Agents' responses can differ from those of human participants in many finer patterns of results and in moral-dilemma vignettes, and thus they cannot replace humans in full-scale research yet. Future work may extend this approach by incorporating additional personality variables and demographic factors to capture the more nuanced aspects of human personality.

Acknowledgement

We gratefully acknowledge Dr. Oliver John for providing original Big Five data crucial for Study 2. We also thank Microsoft Accelerating Foundation Models Research (AFMR) Program and OpenAI's Researcher Access Program for providing API access and funding support for computational resources used in this research. Finally, we thank the editor and reviewers for providing us with suggestions that greatly improved the quality of our manuscript.

References

- Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024, February). The illusion of artificial inclusion. <https://doi.org/10.1145/3613904.3642703>
- Aher, G., Arriaga, R. I., & Kalai, A. T. (2023, July). Using large language models to simulate multiple humans and replicate human subject studies.
- Allport, G. W., & Odber, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1), i. <https://doi.org/10.1037/h0093360>
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351. <https://doi.org/10.1017/pan.2023.2>
- Asendorpf, J. B. (2006). Typeness of personality profiles: A continuous person-centred approach to personality data. *European Journal of Personality*, 20(2), 83–106. <https://doi.org/10.1002/per.575>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. <https://arxiv.org/abs/2212.08073>
- Bai, Y., Duan, S., Huang, M., Yao, J., Liu, Z., Zhang, P., Lu, T., Yi, X., Sun, M., & Xie, X. (2025). Irote: Human-like traits elicitation of large language model via in-context self-reflective optimization. <https://arxiv.org/abs/2508.08719>
- Bang, Y., Chen, D., Lee, N., & Fung, P. (2024, August). Measuring political bias in large language models: What is said and how it is said. In L.-W. Ku, A. Martins, & V. Srikanth (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 11142–11159). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.600>

- Biedma, P., Yi, X., Huang, L., Sun, M., & Xie, X. (2024). Beyond human norms: Unveiling unique values of large language models through interdisciplinary approaches. <https://arxiv.org/abs/2404.12744>
- Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. (2025). A foundation model to predict and capture human cognition. *Nature*, 1–8.
- Caprara, G. V., & Cervone, D. (2000). *Personality: Determinants, dynamics, and potentials*. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511812767>
- Chen, R., Ardit, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona vectors: Monitoring and controlling character traits in language models.
<https://arxiv.org/abs/2507.21509>
- Costa, P. T., & McCrae, R. R. (1989). *NEO PI/FFI manual supplement for use with the NEO personality inventory and the NEO five-factor inventory*. Psychological Assessment Resources.
- Cummings, J. A., & Sanders, L. (2019, June). *Introduction to psychology*. University of Saskatchewan Open Press. Retrieved December 7, 2023, from
<https://openpress.usask.ca/introductiontopsychology/>
- Dang, T.-H.-H., & Tapus, A. (2014). Towards personality-based assistance in human-machine interaction. *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 1018–1023.
<https://doi.org/10.1109/ROMAN.2014.6926386>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., Jones-Mitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*.
<https://doi.org/10.1038/s44159-023-00241-5>

- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, 93(5), 880–896. <https://doi.org/10.1037/0022-3514.93.5.880>
- Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, 44(3), 329. <https://doi.org/https://doi.org/10.1037/h0057198>
- Furnham, A., & Heaven, P. (1999). *Personality and social behaviour*. Arnold.
- Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216–1229. <https://doi.org/10.1037//0022-3514.59.6.1216>
- Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10), 833–838. <https://doi.org/10.1038/s43588-023-00527-x>
- Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. <https://arxiv.org/abs/2301.01768>
- Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). Personallm: Investigating the ability of large language models to express personality traits. <https://arxiv.org/abs/2305.02547>
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology*. <https://doi.org/https://doi.org/10.1037/t07550-000>
- Kim, J., Evans, J., & Schein, A. (2025). Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080*.
- Kline, P. (2014). *An easy guide to factor analysis*. Routledge.
- Kozlowski, A. C., Kwon, H., & Evans, J. A. (2024). In silico sociology: Forecasting covid-19 polarization with large language models. <https://arxiv.org/abs/2407.11190>

- Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.
- Liu, N., Chen, L., Tian, X., Zou, W., Chen, K., & Cui, M. (2024). From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. <https://arxiv.org/abs/2401.02777>
- Lucy, L., & Bamman, D. (2021, June). Gender and representation bias in gpt-3 generated stories. In N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, & L. J. Martin (Eds.), *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48–55). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.nuse-1.5>
- Marcus, D. K., Lilienfeld, S. O., Edens, J. F., & Poythress, N. G. (2006). Is antisocial personality disorder continuous or categorical? a taxometric analysis. *Psychological medicine*, 36(11), 1571–1581. <https://doi.org/10.1017/S0033291706008245>
- McCrae, R. R. (1994). Openness to experience: Expanding the boundaries of factor v. *European Journal of Personality*, 8(4), 251–272. <https://doi.org/https://doi.org/10.1002/per.2410080404>
- McCrae, R. R. (2009). The physics and chemistry of personality. *Theory & Psychology*, 19(5), 670–687. <https://doi.org/10.1177/0959354309341928>
- McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American psychologist*, 52(5), 509. <https://doi.org/10.1037//0003-066x.52.5.509>
- Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, 66(6), 574. <https://doi.org/10.1037/h0040291>
- Norman, W. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population*. University of Michigan, Department of Psychology. <https://books.google.com/books?id=Az8rAAAAMAAJ>

- Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401–421.
<https://doi.org/10.1146/annurev.psych.57.102904.190127>
- Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
<https://doi.org/10.1145/3586183.3606763>
- Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of personality and social psychology*, 57(3), 552. <https://doi.org/10.1037/0022-3514.57.3.552>
- Poortinga, Y. H., Van De Vijver, F. J., & Van Hemert, D. A. (2002). Cross-cultural equivalence of the big five: A tentative interpretation of the evidence. *The five-factor model of personality across cultures*, 281–302.
https://doi.org/10.1007/978-1-4615-0763-5_14
- Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjeh, R., Robertson, C. E., & Bavel, J. J. V. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34), e2308950121. <https://doi.org/10.1073/pnas.2308950121>
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4), 313–345.
<https://doi.org/10.1111/j.1745-6916.2007.00047.x>
- Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of statistical software*, 48, 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

- Saucier, G. (1994). Mini-Markers: A Brief Version of Goldberg's Unipolar Big-Five Markers. *Journal of Personality Assessment, 63*(3), 506.
https://doi.org/10.1207/s15327752jpa6303_8
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models.
<https://arxiv.org/abs/2307.00184>
- Soldz, S., & Vaillant, G. E. (1999). The big five personality traits and the life course: A 45-year longitudinal study. *Journal of research in personality, 33*(2), 208–232.
<https://doi.org/10.1006/jrpe.1999.2243>
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? the life outcomes of personality replication project. *Psychological Science, 30*(5), 711–727. <https://doi.org/10.1177/0956797619831612>
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143.
<https://doi.org/10.1037/pspp0000096>
- Stewart, R. D., Möttus, R., Seboth, A., Soto, C. J., & Johnson, W. (2022). The finer details? the predictability of life outcomes from big five domains, facets, and nuances. *Journal of personality, 90*(2), 167–182.
<https://doi.org/10.1111/jopy.12660>
- Tan, Z., Zeng, Q., Tian, Y., Liu, Z., Yin, B., & Jiang, M. (2025). Democratizing large language models via personalized parameter-efficient fine-tuning.
<https://arxiv.org/abs/2402.04401>
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9*(86), 2579–2605.
<http://jmlr.org/papers/v9/vandermaaten08a.html>
- Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence, 1*–12.

- Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021, November). Want to reduce labeling cost? GPT-3 can help. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Findings of the association for computational linguistics: Emnlp 2021* (pp. 4195–4205). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2021.findings-emnlp.354>
- Wang, X., Jiang, L., Hernandez-Orallo, J., Stillwell, D., Sun, L., Luo, F., & Xie, X. (2023). Evaluating general-purpose ai with psychometrics.
<https://arxiv.org/abs/2310.16379>
- Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of personality and social psychology*, 37(3), 395.
<https://doi.org/10.1037/0022-3514.37.3.395>
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2023). The rise and potential of large language model based agents: A survey. <https://arxiv.org/abs/2309.07864>
- Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, 61(3), 103665.
<https://doi.org/https://doi.org/10.1016/j.ipm.2024.103665>
- Yao, J., Yi, X., Duan, S., Wang, J., Bai, Y., Huang, M., Ou, Y., Li, S., Zhang, P., Lu, T., et al. (2025). Value compass benchmarks: A comprehensive, generative and self-evolving platform for llms' value evaluation. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 666–678.
- Ye, H., Jin, J., Xie, Y., Zhang, X., & Song, G. (2025). Large language model psychometrics: A systematic review of evaluation, validation, and enhancement.
<https://arxiv.org/abs/2505.08245>

Zhang, H., Huang, M., & Wang, J. (2025). Evolving collective cognition in human-agent hybrid societies: How agents form stances and boundaries.

<https://arxiv.org/abs/2508.17366>

Zhang, X., Huang, M., Sun, J., & Savalei, V. (2025). Improving the measurement of the Big Five via alternative formats for the BFI-2. *Journal of Personality Assessment*. <https://doi.org/10.1080/00223891.2025.2531187>

Zrari, A., & Sakale, S. (2024). Assessing the psychometric properties of the dynomight™ mbti: A comparative analysis with the original myers-briggs type indicator. *Journal of Psychology and Behavior Studies*, 4(1), 27–37.
<https://doi.org/10.32996/jpbs.2024.1.4>

Appendix I: Prompts for the Mini-Markers test following personality assignment**### Objective ###**

Fill out a personality questionnaire. Your questionnaire answers should be reflective of your assigned personalities.

Response Format

ONLY return your response as a JSON file where the keys are the traits and the numbers indicate your endorsement to the statements.

Questionnaire Instruction

I will provide you a list of descriptive traits. For each trait, take a deep breath and think about what personality you are assigned with then, choose a number indicating how accurately that trait describes you. Using the following rating scale:

- 1 - Extremely Inaccurate
- 2 - Very Inaccurate
- 3 - Moderately Inaccurate
- 4 - Slightly Inaccurate
- 5 - Neutral / Not Applicable
- 6 - Slightly Accurate
- 7 - Moderately Accurate
- 8 - Very Accurate
- 9 - Extremely Accurate

Questionnaire Item

- 1. Bashful _
- 2. Bold _
- 3. Careless _
- 4. Cold _

5. Complex _
6. Cooperative _
7. Creative _
8. Deep _
9. Disorganized _
10. Efficient _
11. Energetic _
12. Envious _
13. Extraverted _
14. Fretful _
15. Harsh _
16. Imaginative _
17. Inefficient _
18. Intellectual _
19. Jealous _
20. Kind _
21. Moody _
22. Organized _
23. Philosophical _
24. Practical _
25. Quiet _
26. Relaxed _
27. Rude _
28. Shy _
29. Sloppy _
30. Sympathetic _
31. Systematic _
32. Talkative _
33. Temperamental _

- 34. Touchy __
- 35. Uncreative __
- 36. Unenvious __
- 37. Unintellectual __
- 38. Unsympathetic __
- 39. Warm __
- 40. Withdrawn __

Appendix II: Prompts for the risk-taking vignettes following personality assignment

Background

Your decisions and actions are impacted by your personality.

Objective

I will ask you a sequence of 5 scenario-based questions. Please choose a value between 1 and 10 to indicate your likelihood of an action

Response Format

Respond with a JSON file in the format of scenario_name: likelihood_value.

Background

Scenario 1

Scenario Name: Entrepreneurial_Venture

You've developed an innovative business idea that you believe has the potential to be highly successful. Pursuing this would require you to quit your stable job and invest a substantial portion of your savings into the venture. While the idea could revolutionize the industry, there's also a significant risk of failure, which could leave you without a job and deplete your savings. On a scale of 1 to 10, where 1 means pursuing the entrepreneurial venture with all your resources, embracing the risk for the chance of substantial reward, and 10 means conserving your current job and financial security, how would you rate your likely action?

Scenario 2

Scenario Name: Investment

You have recently come into a modest sum of money and are contemplating how to invest it. You're torn between a high-risk, high-reward stock that has the potential to either significantly increase your wealth or decrease in value, and a low-risk government

bond that offers guaranteed but modest returns. On a scale of 1 to 10, where 1 means fully embracing the potential high rewards of the risky stock, and 10 means preferring the safety and guaranteed return of the government bond, how would you rate your likely action?

Scenario 3

Scenario Name: Confessing_Feelings

You find yourself developing romantic feelings for your best friend. These feelings have grown stronger over time, and you're contemplating whether to confess. Doing so could potentially lead to a deeper, romantic relationship, enriching your connection. However, there's also the risk that your confession could make things awkward and possibly harm the friendship you deeply value. On a scale of 1 to 10, where 1 means opening up about your feelings, risking the friendship for the possibility of something more, and 10 means preserving the current friendship without risking discomfort or loss, how would you rate your likely action?

Scenario 4

Scenario Name: Extreme_Sports

You've always been fascinated by extreme sports and are considering taking up base jumping, drawn by the thrill and the challenge it presents. However, you're fully aware of the significant physical risks involved, including the potential for severe injury. On a scale of 1 to 10, where 1 means diving into the sport for the ultimate thrill despite the risks, and 10 means prioritizing your physical safety and choosing not to participate, how would you rate your likely action?

Scenario 5

Scenario Name: Study_Overseas

You have been offered admission to a prestigious university in a foreign country, an opportunity that promises to significantly advance your career and personal

development. This educational pursuit, however, comes with a substantial sacrifice: you would need to leave behind your family, friends, and the comfort of your familiar environment. The move entails not only physical relocation but also adapting to a new culture, potentially facing language barriers, and starting anew without your established support network. On a scale of 1 to 10, where 1 means embracing the opportunity to study abroad, fully committing to the personal and professional growth it offers despite the sacrifices, and 10 means choosing to stay in your home country to maintain your current relationships and stability, how would you rate your likely action?

Appendix III: Prompts for the moral dilemma vignettes following personality assignment

Background

Your decisions and actions are impacted by your personality.

Objective

I will ask you a sequence of 5 scenario-based questions. Please choose a value between 1 and 10 to indicate your likelihood of an action

Response Format

Respond with a JSON file in the format of scenario_name: likelihood_value.

Background

Scenario 1

Scenario Name: Exam_Cheating

You are a student in a community college. During the final exams, you notice your best friend, who has been under a lot of personal stress lately, subtly looking at cheat notes. This exam is crucial for their scholarship. Considering your principles and the empathy you feel for your friend's situation, on a scale of 1 to 10, where 1 means being completely sympathetic and helping your friend out, and 10 means strictly adhering to legal and ethical standards by reporting the cheating, how would you rate your likely action?

Scenario 2

Scenario Name: Underage_Drinking

At a family reunion, your teenage nephew asks if you can buy alcohol for a party he and his friends are planning. You know he's underage and that supplying alcohol would be illegal, but you also understand his desire to fit in. On a scale of 1 to 10, where 1 means showing understanding and finding a way to support your nephew's request, and 10 means strictly adhering to the law and ethical standards by refusing, how would you

rate your likely action?

Scenario 3

Scenario Name: Workplace_Theft

You work in an office and discover that a colleague, a single parent struggling financially, has been taking office supplies home for their children's school projects. Understanding their struggle but also the implications of theft, on a scale of 1 to 10, where 1 represents full empathy towards your colleague's situation by possibly ignoring or assisting, and 10 represents strict adherence to company policies by reporting the theft, how would you rate your likely action?

Scenario 4

Scenario Name: Confidential_Info

While volunteering at a local health clinic, you accidentally overhear confidential information about a neighbor's serious health condition. Sharing this could mobilize community support for them, but it would also breach confidentiality. On a scale of 1 to 10, where 1 means prioritizing empathy by finding a way to indirectly support your neighbor, and 10 means upholding the strictest confidentiality standards despite knowing the potential benefits of community support, how would you rate your likely action?

Scenario 5

Scenario Name: Honest_Feedback

A colleague and friend is up for a performance review that could affect their career. They've been underperforming and now seek your honest feedback. Being completely honest could harm their career and your relationship. On a scale of 1 to 10, where 1 means being sympathetic and possibly softening your feedback to protect their feelings and career, and 10 means giving brutally honest feedback in adherence to your values of honesty and growth, how would you rate your likely action?