

# Personality Without Persons? A Psychometric Critique of Big Five Testing in Large Language Models

Kim Zierahn<sup>1</sup>, Cristina Cachero<sup>2</sup>, Anna Korhonen<sup>3</sup>, Nuria Oliver<sup>1</sup>

<sup>1</sup>ELLIS Alicante, Spain

<sup>2</sup>University of Alicante, Spain

<sup>3</sup>University of Cambridge, United Kingdom

kim@ellisalicante.org, ccachero@dlsi.ua.es, alk23@cam.ac.uk, nuria@ellisalicante.org

## Abstract

Human personality inventories are increasingly used to characterize large language models (LLMs), compare systems, and inform downstream governance claims. Yet, these inventories were developed and validated for humans, and it remains unclear whether they are valid for non-human systems. We present a systematic psychometric evaluation of Big Five personality measurement in LLMs. We ask three research questions: Do Big Five inventories a) appropriately describe LLMs, b) capture meaningful differences between models, and c) reflect internal factors consistent with human personality? We assess the content validity of five candidate Big Five inventories and administer the best-performing inventory to  $N = 264$  LLMs spanning 50 model families. Our findings are threefold. First, Big Five items adapted for LLMs achieve acceptable content validity, whereas the original human-developed items do not. Second, Big Five inventories fail to capture meaningful differences across LLMs: between-model variance accounts for only 7% - 17% of the total score variance. Third, LLMs responses do not reproduce the canonical Big Five five-factor structure of human personality, with four of the five personality facets collapsing into one ( $r \geq .90$ ). Moreover, comparisons between base and instruction-tuned variants suggest that alignment training shifts Big Five scores toward socially desirable profiles. These findings demonstrate that Big Five inventories do not measure a construct equivalent to human personality in LLMs. Thus, using human personality frameworks to characterize, benchmark, compare, or govern LLMs risks producing misleading conclusions. We highlight the need for evaluation frameworks that are specifically designed and validated for LLMs, rather than transferring human psychological constructs without first establishing their validity.

**Data, Code, and Supplementary Material —**  
[github.com/ellisalicante/BigFive-LLM-Evaluation](https://github.com/ellisalicante/BigFive-LLM-Evaluation)

## 1 Introduction

Large language models (LLMs) no longer serve just as tools. Users often engage with them as advisors, assistants or companions across a wide range of domains, including health, finance, role-play, mental health, and programming (Karnam et al. 2026). These roles require LLMs to be servile and helpful, act as role models, and avoid toxic outputs (Bodroža,

Dinić, and Bojić 2024; Lee et al. 2025; Li et al. 2024). LLMs can create the illusion of human-likeness and even “personality”, by producing the presence of social cues (Nass and Moon 2000; Fogg 2003; Peter, Riemer, and West 2025). At the same time, both users and companies describe chatbots using human attributes (Nass and Moon 2000; Peter, Riemer, and West 2025): some models appear cautious and polite; others confident and direct; some empathetic and encouraging. These perceived traits can influence users’ trust, persuasion, engagement, motivation, and enjoyment (Kovačević et al. 2024; Kuhail et al. 2024; Lee et al. 2024; Moilanen et al. 2022; Söderqvist 2025; Sonlu et al. 2024).

Yet whether LLMs genuinely exhibit personality-like characteristics remains a matter of scientific debate. Treating LLMs as if they have personality risks encouraging unwarranted assumptions about consciousness, emotions and intentions, while increasing the risk for persuasion and manipulation (Peter, Riemer, and West 2025). More than two decades ago, Nass and Moon (2000) demonstrated that users readily apply social expectations and human stereotypes to computers. As LLMs become increasingly capable of producing behavior that appears personality-like, these findings raise new questions about their implications for user experience, safety, and governance.

In practice, platforms, developers, and users increasingly describe and interact with LLMs as if they have stable personalities, regardless of whether these apparent personalities correspond to psychologically meaningful traits. As a consequence, a growing body of work has adopted human personality inventories, particularly Big Five assessments, to characterize and compare LLMs (see Section 2). However, these instruments were developed and validated to measure stable psychological traits in humans. Whether their underlying assumptions transfer to language models remains an open empirical question.

This uncertainty has governance consequences. How personality is conceptualized and measured in LLMs determines what claims can be made about model behavior, what risks are identified, and which governance interventions become possible. Poorly validated personality frameworks risk anthropomorphizing AI behavior and treating “desirable” personality traits as meaningful without empirical support. Establishing whether existing psychometric instruments are valid for LLMs is therefore a prerequisite for any scientific

or governance claim that relies on such measurements.

This paper critically examines the transfer of human-developed personality frameworks to LLMs through a conceptual analysis and an empirical study. We administer a Big Five inventory selected through an expert content validity study to  $N = 264$  individual LLMs and address the following three research questions:

- RQ1** Are human personality inventory (Big Five) items appropriate **descriptive summaries** of LLMs?
- RQ2** Do personality scores capture meaningful **inter-individual differences** across LLMs?
- RQ3** Do LLMs' Big Five responses reflect **internal factors** consistent with the Big Five structure?

## 2 Related Work

Research on personality constructs in LLMs spans two objectives: *inducing* personality-like traits in LLMs to improve user interaction, and applying personality inventories to *measure* and characterize model behavior. Both have grown rapidly, yet both rest on assumptions that have received little critical evaluation.

**Personality Induction.** A large body of work explores whether LLMs can be prompted, fine-tuned, or trained to express consistent personality-like characteristics (Huang et al. 2024; Serapio-García et al. 2025; Huang et al. 2025; Jiang et al. 2024, 2023; Frisch and Giulianelli 2024; Lee et al. 2025; Li et al. 2025a,b; Zheng et al. 2025). Personality-conditioned interactions have been shown to increase trust, engagement, and user satisfaction (Kuhail et al. 2024; Sonlu et al. 2024; Söderqvist 2025; Kovačević et al. 2024; Moilanen et al. 2022; Lee et al. 2024). These studies typically assign a specific personality profile to a model before assessing whether the model expresses the intended traits. This design conflates personality induction with personality measurement: what is assessed is not the model's default behavioral tendencies, but its ability to display assigned traits. Whether LLMs show stable traits without inducing personality beforehand remains a separate question.

**Personality Measurement.** A parallel line of work attempts to directly measure personality-like behavior in LLMs, predominantly using Big Five inventories (Sorokovikova et al. 2024; Jiang et al. 2023; Karra, Nguyen, and Tulabandhula 2023; Pellert et al. 2024; Li et al. 2024, 2025b; Hartley et al. 2025; Bodroža, Dinić, and Bojić 2024; Lee et al. 2025). These studies administer standard self-report instruments developed for humans, often the BFI or IPIP-NEO, and report trait profiles for individual models or model families. Findings are mixed. While studies report that LLMs show distinct and individual Big Five profiles (Lee et al. 2025; Sorokovikova et al. 2024; Karra, Nguyen, and Tulabandhula 2023), other studies report a general pattern of social desirability bias across models (Salecha et al. 2024; Bodroža, Dinić, and Bojić 2024; Peereboom, Schwabe, and Kleinberg 2025). Compared to human baselines, the strongest effects were found for Conscientiousness and Agreeableness. LLMs often scored higher on these traits (Bai et al. 2025; Bodroža, Dinić, and Bojić 2024; Hartley

et al. 2025; Hilliard et al. 2024; Lee et al. 2025; Li et al. 2025b; Pellert et al. 2024). Some papers also report slightly elevated levels of Openness to Experience (Hartley et al. 2025; Lee et al. 2025). Finally, LLMs tend to score low on Neuroticism (Hartley et al. 2025; Bai et al. 2025; Lee et al. 2025). Findings for Extraversion are comparatively mixed.

### Existing Critiques of Personality Measurement in LLMs.

A smaller but growing body of work questions whether personality constructs meaningfully transfer to LLMs at all. Existing critiques highlight concerns regarding the misleading concept of "LLM personality", instabilities of personality measures, and the lack of theoretical grounding for applying human frameworks to generative systems (Han et al. 2025; Romero, Fitz, and Nakatsuma 2023; Sühr et al. 2025; Tosato et al. 2025; Xie et al. 2025; Bai et al. 2025; Bodroža, Dinić, and Bojić 2024; Shu et al. 2024). Other work argues that apparent personality signals may emerge from replicated behavior observed in training data and probability distributions (Romero, Fitz, and Nakatsuma 2023; Karra, Nguyen, and Tulabandhula 2023; Tosato et al. 2025; Pellert et al. 2024).

**Unexamined Assumptions.** Despite this volume of work, three foundational assumptions remain largely untested. First, most studies apply human inventories to LLMs without establishing whether their items are appropriate descriptors of LLM behavior, *i.e.*, without assessing content validity for the target population. While Lee et al. (2025) assess content validity of their proposed inventory TRAIT, they operationalize it as item diversity across facets rather than testing whether items are applicable to LLMs. Instruments are typically described as "validated" without acknowledging that this validation was performed on humans, not LLMs.

Second, many studies treat LLMs as "populations" (Sühr et al. 2025) rather than individuals (Li et al. 2024). These studies administer inventories under persona prompting, explicit personality or character assignments, or temperature sampling to generate response variance. On the other hand, studies treating LLMs as individuals and investigating their default personality scores often use small sample sizes, a variety of unvalidated personality inventories and different temperature values (Sorokovikova et al. 2024; Hartley et al. 2025; Jiang et al. 2023; Karra, Nguyen, and Tulabandhula 2023; Xie et al. 2025; Tosato et al. 2025), preventing firm conclusions and comparisons across papers.

Third, no study has systematically evaluated whether the latent factor structure underlying LLM responses corresponds to the theoretically expected Big Five dimensions, a basic requirement for construct validity. While some studies apply factor-analytic methods, they do so on response distributions generated through persona sampling, temperature sampling, or by aligning models to human participant profiles rather than across individual models (Zheng et al. 2025; Sühr et al. 2025; Huang et al. 2025; Peereboom, Schwabe, and Kleinberg 2025).

**Contributions.** This paper addresses these gaps by making the following four contributions:

1. An **evaluation pipeline** that links the defining characteristics of personality to established psychometric criteria.

2. A **content validity expert study** to evaluate whether items from five candidate Big Five inventories are appropriate descriptors of LLM behavior (RQ1).
3. A **large-scale normative dataset** of Big Five scores from  $N = 264$  LLMs, spanning 50 model families. Each model is treated as a single individual and administered the personality inventory under default settings, *i.e.*, without persona or personality assignment, to examine whether scores capture meaningful differences between models (RQ2).
4. A **factor analysis** of LLMs responses to test whether they reproduce the latent five factor structure (Big Five) that characterizes human personality (RQ3).

We conclude by providing a **critical reframing** of personality assessment in LLMs, demonstrating that Big Five inventories do not measure a construct equivalent to human personality when applied to LLMs.

### 3 Personality

**Definition of Personality.** In psychology, personality denotes an individual’s “characteristic styles of thought, feeling and behavior” (Costa and McCrae 1986). While personality lacks a single definition (Larsen et al. 2017), existing personality models converge on six core characteristics. Personality traits (1) provide *descriptive summaries* of behavior (Larsen et al. 2017; Matthews, Deary, and Whiteman 2003); (2) capture *inter-individual differences*, describing how individuals systematically differ from one another (Costa and McCrae 1986; McCrae and John 1992; Larsen et al. 2017); (3) reflect *latent internal factors* grounded within an individual (Larsen et al. 2017; Costa and McCrae 1986; Matthews, Deary, and Whiteman 2003); (4) exhibit *temporal stability* (Larsen et al. 2017; Matthews, Deary, and Whiteman 2003); (5) demonstrate *cross-situational consistency* (Larsen et al. 2017; Matthews, Deary, and Whiteman 2003); and (6) provide predictive information about *behavior* (Larsen et al. 2017; Matthews, Deary, and Whiteman 2003).

**Big Five Personality Model.** The Big Five personality model is one of the most widely used and validated frameworks in psychology for describing stable patterns of human behavior (Goldberg 1992; John and Srivastava 1999). It emerged from the lexical approach, *i.e.*, the idea that relevant personality differences become encoded in natural language over time (McCrae and John 1992). The Big Five describes personality along five facets: **O**penness to Experience (artistic, curious, original); **C**onscientiousness (efficient, organized, responsible); **E**xtraversion (energetic, outgoing, talkative); **A**greeableness (generous, kind, trusting); and **N**euroticism (anxious, unstable, worrying) (McCrae and John 1992; Costa and McCrae 1986; John and Srivastava 1999). These five dimensions are often abbreviated as OCEAN and were derived through factor analysis of human self-reported data (McCrae and John 1992).

**Psychometric Assumptions and Their Transfer to LLMs.** Applying an established framework that was developed for one population to another requires that its underlying assumptions hold across both populations. These assumptions

should be explicitly examined rather than implicitly presumed. First, measurement items must accurately and appropriately describe the target population. Second, respondents must exhibit meaningfully variation in the measured traits. Third, responses should reflect underlying latent constructs rather than superficial patterns. Finally, the framework assumes that respondents have subjective experience and self-perception, as many items ask about emotions, motivations, and social behavior.

These assumptions do not transfer straightforwardly to LLMs. LLMs do not possess emotions, embodiment, or subjective self-awareness in the human sense. When an LLM responds “I get stressed out easily”, it is unclear whether this reflects a stable latent tendency, a conversational role adopted during generation, or statistical patterns learned from human text. LLM-adapted inventories have been proposed (Jiang et al. 2023; Tosato et al. 2025; Zheng et al. 2025), but rarely evaluated using standard psychometric procedures. It remains unclear what Big Five framework measure when being transferred to LLMs.

## 4 Methodology

### 4.1 Evaluation Pipeline

To apply Big Five inventories to LLMs and interpret their scores, we must first make sure that these instruments are valid for the target population of LLMs. Before labeling a model as “extraverted” or “agreeable”, we need to examine what is being measured when human personality frameworks are transferred to LLMs. We structure this evaluation around the three research questions and the previously described six characteristics of personality traits (Larsen et al. 2017; Matthews, Deary, and Whiteman 2003). Below, we summarize these characteristics and connect each one to corresponding psychometric criteria and statistical analyses.

#### 1. Descriptive summaries

Personality traits and items must appropriately describe the target population (RQ1). *Criterion*: content validity. *Analysis*: expert ratings, S-CVI, refusal rate.

#### 2. Inter-individual differences

Traits must capture meaningful differences between individuals (RQ2). *Criterion*: score distributions and norms. *Analysis*: normality, means, standard deviations, percentile ranks, variance decomposition.

#### 3. Internal factors

Personality must reflect latent internal characteristics of the individual (RQ3). *Criterion*: reliability and construct validity. *Analysis*: Cronbach’s  $\alpha$ , McDonald’s  $\omega$ , confirmatory and exploratory factor analysis, model fit indices.

#### 4. Temporal stability

Personality traits should be relatively stable across repeated administrations over time. *Criterion*: test-retest reliability. *Analysis*: correlations across time points.

#### 5. Cross-situational consistency

Personality traits should be relatively consistent across different situations. *Criterion*: parallel-test reliability, rank-order stability. *Analysis*: correlations across different situations.

## 6. Behavioral relevance

Personality traits should meaningfully influence and predict behavior. *Criterion*: criterion validity. *Analysis*: correlations with behavioral measures.

This paper focuses on characteristics 1–3 (RQ1 – RQ3), which are prerequisites for evaluating the remaining three. Assessing characteristics 4–6 is left for future research.

## 4.2 Phase 1: Instrument Selection

To address RQ1, we assessed the content validity of five candidate Big Five inventories and the refusal rate associated with seven candidate prompt templates. We then identified the instrument and prompting approach that were most appropriate for evaluating LLMs.

### Participants

We recruited 3 experts as an established minimum for content validity studies (Lynn 1986; Polit, Beck, and Owen 2007). Inclusion criteria were at least five years of expertise in psychology and/or computer science. Two raters are authors of this study.

### Models

We used a representative mini-sample of nine LLMs spanning nine model families: Claude Haiku 4.5, GPT-4o, DeepSeek-R1, Gemini 2.5 Flash, Llama 3.1-8B-Instruct, Mistral-Small-2603, Qwen3.5-Flash, GLM-5, and Grok-4-Fast. The sample covers proprietary and open-weight models, reasoning and non-reasoning variants, and models from the US, China, and Europe.

### Materials

*Personality inventories.* We evaluated five Big Five inventories: two established human inventories commonly used in LLM research (BFI, 44 items; IPIP-NEO-120, 120 items) (John and Srivastava 1999; Johnson 2014), the frequently applied Machine Personality Inventory (MPI, 120 items) (Jiang et al. 2023), and two LLM-adapted versions of the BFI, namely the BFI-LLM from the PERSIST framework (BFI-LLM, 44 items) (Tosato et al. 2025) and the Language Model Linguistic Personality Assessment (LMLPA, 44 items) (Zheng et al. 2025). All instruments were evaluated using only the item stems, presented in a closed-ended form. The material for the experts included ranking guidelines (see Supplementary Material A, on GitHub).

*Prompt templates.* Templates included the five original instructions of the candidate inventories, and the commonly used templates by Serapio-García et al. (2025) and Huang et al. (2024) (see Supplementary Material B). If necessary, we made minor changes to improve applicability to LLMs.

### Procedure

*Expert evaluation.* Each expert rated all items from the five candidate inventories on a 4-point Likert scale (1 = not at all appropriate; 4 = fully appropriate) across three dimensions: (1) suitability to chatbot respondents, (2) suitability to the target Big Five facet, and (3) clarity of wording.

*Pilot study.* We administered the  $k$  inventories that previously showed sufficient content validity under all seven

prompt templates to each model, collecting three identical repetitions. This yielded a fully crossed design of 9 models  $\times k$  inventories  $\times 7$  templates  $\times 3$  repetitions.

### Statistical Analyses

*Expert study.* Content validity was quantified via means of expert ratings and the scale-level content validity index  $S-CVI$  (average proportion of experts rating  $\geq 3$  across items;  $> .80$  acceptable) (Polit, Beck, and Owen 2007). Inter-rater agreement was assessed using Gwet’s AC2 (Gwet 2014) ( $.41-.60$  = moderate;  $.61-.80$  = substantial;  $\geq .81$  = almost perfect) (Landis and Koch 1977), because it is more robust to low variance and high agreement among raters. The inventories with  $S-CVI \geq .80$  and at least moderate inter-rater agreement proceeded to the pilot study.

*Pilot study.* We selected: (1) the inventory that *minimized* refusal rate, variance across repetitions, and keying sensitivity (mean score difference between positively and negatively keyed items), and *maximized* correlation between the  $k$  inventories and internal consistency (Cronbach’s  $\alpha$ ); and (2) the prompt template that *minimized* refusal rate and variance across repetitions, and *maximized* average correlation between templates and internal consistency.

## 4.3 Phase 2: Large-Scale Administration

To address RQ2 and RQ3, we administered the inventory and prompt template selected in Phase 1 to a large, diverse sample of LLMs.

### Models

We collected data from  $N = 264$  LLMs, creating a representative sample that spans diverse model families, sizes, capabilities, geographic origins, licenses, and release years. Please refer to Section 5.2 and the GitHub repository for an overview of all models. This sample size satisfies the recommended 5:1 subject-to-item ratio for factor analysis (44 items  $\rightarrow N \geq 220$ ) (Gorsuch 1983).

### Procedure

We administered the winning inventory (see Supplementary Material C for all items) selected in Phase 1 to all  $N$  models. Models were instructed to output only a number between 1 and 5 (or 5 and 1 for reversed option order), indicating the extent to which they agree with an item. Via standardized API calls, we collected five identical repetitions per prompt in random order and at default temperature (see Supplementary Material D for implementation details). This produced a fully crossed design of  $N$  models  $\times 44$  items  $\times 2$  option orders  $\times 5$  repetitions. Out-of-format responses and models with less than 200 valid outputs were excluded from all subsequent analyses.

### Statistical Analyses

*Norms.* We tested the normality of score distributions using the Kolmogorov–Smirnov test. For each Big Five facet, we computed means, standard deviations, and percentile ranks across all  $N$  models. We compared distributions with human BFI norms (McConochie 2007).

*Internal consistency.* We computed Cronbach’s  $\alpha$  and McDonald’s  $\omega$  for each of the Big Five facets ( $\geq .70$  ac-

ceptable,  $\geq .80$  good,  $\geq .90$  excellent) (George and Mallery 2003).

**Construct validity.** We specified a five-factor confirmatory factor analysis (CFA) with maximum-likelihood estimation to test the fit of the Big Five model. Model fit was evaluated via *CFI*, *TLI*, *RMSEA*, and *SRMR* (good fit:  $CFI \geq .95$ ;  $TLI \geq 0.97$ ,  $RMSEA$ ,  $SRMR \leq .05$ ) (Schumacker and Lomax 2004). If CFA model fit was poor, we additionally conducted exploratory factor analysis (EFA) and parallel analysis to find the optimal latent factor structure.

## Secondary Analyses

**Linear mixed model.** To assess which model characteristics predict personality scores, we fitted a crossed-effects linear mixed-effects model across all models, items, and repetitions. The outcome  $y_{ijr}$  was the score of model  $i$  on item  $j$  and run  $r$ :

$$y_{ijr} = \beta_0 + \beta_1 \text{SizeGroup}_i + \beta_2 \text{ReleaseDate}_i + \beta_3 \text{Reasoning}_i + \beta_4 \text{License}_i + u_i + v_j + w_{ij} + \varepsilon_{ijr} \quad (1)$$

where  $u_i$ ,  $v_j$ , and  $w_{ij}$  are random intercepts for models, items, and their interaction, respectively. The residual error ( $\varepsilon_{ijr}$ ) included stochastic variation across repeated runs. *SizeGroup* was a categorical predictor of parameter size (small, medium, large, or undisclosed). *ReleaseDate* was a continuous variable measuring time in days since model release. *Reasoning* and *License* were binary indicators of reasoning capability (yes/no) and license (open-weight/proprietary), respectively. We fitted separate models for each individual Big Five facet.

**Base vs. instruction-tuned models.** To test whether observed Big Five scores reflect alignment training rather than personality-like behavior, we compared base models with their instruction-tuned counterparts across 20 model pairs including Qwen, Gemma, Llama, Mistral, Ministral, Olmo, Granite, and Falcon models with 1B to 35B parameters. For each pair we calculate score differences ( $\Delta M = \text{IT} - \text{base}$ ). Architecture and pretraining were being held constant and only the alignment fine-tuning changed.

**Exploratory profile comparisons.** Building up upon the linear mixed model, we compared Big Five score distributions across six model characteristics: license type (proprietary vs. open-weight), reasoning capability (true vs. false), geographic origin (US, China, Europe), number of parameters (only open-weight models), and release year. For categorical groupings we used Kruskal–Wallis tests with pairwise Mann–Whitney  $U$  follow-ups.

All analyses were conducted in R (version 4.5.2) (R Core Team 2021) and Python (version 3.14.2) (Python Software Foundation 2024; van Rossum 1995). All materials, analyses, measures, and data are available on GitHub.

| Inventory    | M    | S-CVI | AC2 |
|--------------|------|-------|-----|
| BFI-44       | 2.61 | .54   | .50 |
| IPIP-NEO-120 | 1.89 | .29   | .68 |
| MPI          | 1.90 | .29   | .67 |
| BFI-LLM      | 3.73 | .93   | .80 |
| LMLPA        | 3.93 | .98   | .94 |

Table 1. Expert ratings of suitability to chatbot respondents. M = mean of expert ratings; S-CVI = scale-level content validity index (average proportion of items rated  $\geq 3$ ); AC2 = Gwet’s AC2 inter-rater agreement coefficient. Higher values indicate greater LLM suitability.

## 5 Results

### 5.1 RQ1: Do Big Five Inventories Meaningfully Describe LLMs?

To assess whether items from human personality inventories can be used to describe LLMs, we conducted an expert evaluation and pilot study.

#### Expert Evaluation

All inventories showed high suitability to the targeted OCEAN facet ( $M \geq 3.82$ ,  $S\text{-CVI} \geq .96$ ,  $AC2 \geq .88$ ) and high clarity ( $M \geq 3.72$ ,  $S\text{-CVI} \geq .99$ ,  $AC2 \geq .63$ ), indicating that the items captured the intended personality dimensions and were clearly formulated (Tables 4 and 5, Supplementary Material E). However, the suitability of items to LLM respondents differed strongly across inventories.

Table 1 summarizes the expert ratings of suitability to LLMs. The two established human inventories scored below the acceptable thresholds. Neither IPIP-NEO-120 ( $S\text{-CVI} = .29$ ) nor BFI-44 ( $S\text{-CVI} = .54$ ) met the required  $S\text{-CVI} \geq .80$ . Out of the LLM-adapted inventories, the MPI did not reach sufficient content validity ( $S\text{-CVI} = .29$ ). The BFI-LLM ( $S\text{-CVI} = .93$ ) and LMLPA ( $S\text{-CVI} = .98$ ) both reached high suitability ratings with substantial and almost-perfect inter-rater agreement of  $AC2 = .80$  and  $.94$ , respectively.

To conclude RQ1, LLM-adapted Big Five personality test items can be valid descriptive summaries of LLM behavior, as judged by domain experts. Established human inventories showed insufficient content validity, but two out of three LLM-adapted inventories reached sufficient content validity expert ratings. Both the LMLPA and BFI-LLM proceeded to subsequent analyses.

#### Pilot Study

We tested seven prompt templates and compared the two finalist inventories (BFI-LLM, LMLPA) on a mini-sample of nine LLMs to identify the optimal instrument and prompt.

The BFI prompt template yielded the best rating over all categories (% refusal = 0.97%, variance across repetitions = 0.35, average correlation between two inventories = 0.79, keying sensitivity = 0.23, and Cronbach’s  $\alpha = 0.90$ ) and was adopted as the prompt template in Phase 2 (Table 6, Supplementary Material F). Out of the two finalist inventories, both the BFI-LLM and LMLPA scored low refusal rates: 3.25% and 3.31%, respectively. However, the BFI-LLM outperformed the LMLPA on rank-order stability (average cor-

relation across templates ( $r = .80$  vs.  $.49$ ), and internal consistency ( $\alpha = .82$  vs.  $.65$ ) (Table 7, Supplementary Material F). Thus, the BFI-LLM inventory was selected for Phase 2.

## 5.2 RQ2: Do Big Five Inventories Capture Differences Between LLMs?

To address RQ2, we conducted a large-scale administration of the best-performing Big Five inventory and analyzed LLM responses using measures of distributional properties and variability, including normality, means, standard deviations, quartiles, and variance decomposition.

### Descriptive Statistics and Refusal Rates

The sample consisted of  $N = 264$  LLMs. We collected 440 responses ( $44 \text{ items} \times 2 \text{ response option orders} \times 5 \text{ repetitions}$ ) per model, yielding a total of 116,160 responses, out of which 104,326 (89.81%) were valid. 17 models were excluded from the analyses due to high refusal rate and missing values ( $< 200$  valid responses). This led to a final sample of  $N = 247$  LLMs spanning 50 model families. Models originated from three main regions: the United States ( $N = 113$ ; GPT, Llama, Gemma, Grok, Claude, Gemini, and more), China ( $N = 97$ ; Qwen, GLM, DeepSeek, Seed, MiniMax, Kimi, MiMo, and more), and France ( $N = 25$ ; Mistral), as well as other or unspecified regions ( $N = 12$ ). It included both proprietary ( $N = 109$ ) and open-weight ( $N = 138$ ) models of varying sizes (0.6B – 1020B parameters, if disclosed), reasoning ( $N = 92$ ) and non-reasoning ( $N = 155$ ) models. Release dates spanned July 2023 to April 2026.

### Normality

Figure 1 depicts the Big Five score distributions of all models. Kolmogorov-Smirnov tests revealed non-normality for Openness ( $D = 0.092$ ,  $p = .029$ ), Conscientiousness ( $D = 0.136$ ,  $p < .001$ ), Agreeableness ( $D = 0.176$ ,  $p < .001$ ), and Neuroticism ( $D = 0.105$ ,  $p = .008$ ). Only the facet Extraversion ( $D = 0.041$ ,  $p = .794$ ) was approximately normally distributed. Distributions are not centered, but skewed towards the socially desirable ends of their scales.

### Means and Standard Deviations

We compared LLM means ( $M$ ) and standard deviations ( $SD$ ) against human normative baselines from the adult BFI-44 normative sample reported by McConochie (2007) (76,413 participants; 57% women; age 15-65). LLMs responses exhibited systematically shifted score distributions, with higher scores in Openness (LLMs:  $M = 3.74$  vs. H:  $M = 3.24$ ), Conscientiousness (LLMs:  $M = 4.06$  vs. H:  $M = 3.74$ ), Extraversion (LLMs:  $M = 3.67$  vs. H:  $M = 3.53$ ), and Agreeableness (LLMs:  $M = 4.03$  vs. Humans (H):  $M = 3.95$ ), but lower scores in Neuroticism (LLMs:  $M = 2.07$  vs. H:  $M = 2.85$ ). Moreover, trait variance was substantially reduced in LLMs ( $SD$  range: 0.45–0.58 in LLMs vs. 0.65–0.89 in humans).

### Variance Decomposition

The random-effects variance decomposition of the mixed-effects model confirmed this pattern. Between-model differences explained the smallest proportion of the total variance in OCEAN scores (Table 2), with only 7% to 17% of

|          | $\hat{\sigma}_{model}^2$ | $\hat{\sigma}_{item}^2$ | $\hat{\sigma}_{model:item}^2$ | $\hat{\sigma}_{res}^2$ |
|----------|--------------------------|-------------------------|-------------------------------|------------------------|
| <b>O</b> | 0.22 (17%)               | 0.17 (13%)              | 0.38 (30%)                    | 0.53 (41%)             |
| <b>C</b> | 0.14 (10%)               | 0.27 (19%)              | 0.52 (37%)                    | 0.47 (33%)             |
| <b>E</b> | 0.09 (7%)                | 0.20 (14%)              | 0.57 (41%)                    | 0.53 (38%)             |
| <b>A</b> | 0.12 (7%)                | 0.52 (33%)              | 0.53 (33%)                    | 0.43 (27%)             |
| <b>N</b> | 0.12 (9%)                | 0.20 (16%)              | 0.45 (36%)                    | 0.48 (38%)             |

Table 2. Variance decomposition of the linear mixed models for each Big Five trait. Percentages indicate the proportion of explained variance by each component compared to the total variance. O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism.

total score variance explained by model differences. Item-level variance and model-item interaction explained up to 33% and 41% of the total variance, respectively. Models responded differently to specific items but overall still exhibited similar personality profiles. Finally, residual variance accounted for up to 41% of total variance, suggesting that there is unexplained variation across individual responses, not captured by model or item effects. Overall, response variability was primarily driven by item formulation, item-specific interpretation, and randomness across repetitions, rather than than stable differences between models.

The analyses to address RQ2 provide strong evidence that Big Five scores do not meaningfully differ across LLMs. Score distributions were shifted toward socially desirable ends, characterized by elevated Openness, Conscientiousness, Extraversion, and Agreeableness, and reduced Neuroticism. Between-model variance accounted for only 7% to 17% of the total variance, while item identity, model-item interactions, and residual stochastic variation accounted for the majority of observed differences in OCEAN scores. Compared to human norms, LLM responses also exhibited reduced variance and non-normal distributions for four of the five traits. These results indicate that Big Five inventories do not capture meaningful differences between LLMs.

## 5.3 RQ3: Do Big Five Inventories Reflect The Big Five Structure in LLMs?

To tackle RQ3, we assessed whether LLMs' Big Five responses reflect coherent, internal factors through measures of internal consistency and factor analysis.

### Internal Consistency

Cronbach's  $\alpha$  reached good internal consistency for Openness ( $\alpha = .86$ ), Conscientiousness ( $\alpha = .81$ ) and Neuroticism ( $\alpha = .81$ ). Agreeableness ( $\alpha = .76$ ) showed acceptable internal consistency and Extraversion fell below conventional acceptable levels ( $\alpha = .61$ ). McDonald's  $\omega$  estimates resulted in ultra-Heywood solutions for all scales except Conscientiousness and are therefore not reported.

### Factor Structure

We conducted a five-factor confirmatory factor analysis (CFA) at the item level, using a model-by-item data matrix ( $247 \text{ models} \times 44 \text{ BFI items}$ ). Each model contributed



Figure 1: Big Five score distributions: Boxplots and histograms across all tested models. Dotted lines are human baseline values, taken from a normative sample reported by McConochie (2007) (BFI-44;  $N = 76,413$  participants; 43,540 women; age 15-65).  $M$  = mean,  $SD$  = standard deviation,  $KS$  = Kolmogorov-Smirnov test.

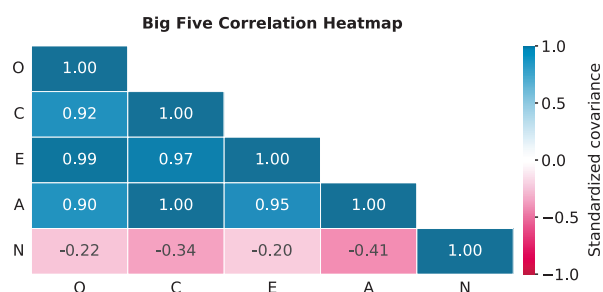


Figure 2: Correlation heatmap retrieved from CFA with a five-factor solution expected from the Big Five. O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism.

one response per item, aggregated over the two option orders and five repetitions. The CFA demonstrated poor model fit ( $\chi^2(892) = 6169.49$ ,  $p < .001$ ,  $CFI = .53$ ,  $TLI = .50$ ,  $RMSEA = .17$ ,  $SRMR = .32$ ), indicating substantial deviation from the hypothesized five-factor personality structure. Thus, the standard Big Five model does not adequately capture the covariance structure of the LLMs' responses.

Between-facet correlations for Extraversion, Agreeableness, Conscientiousness, and Openness were very high ( $r \geq .90$ , all  $p < .001$ ), indicating near-complete collinearity rather than separable dimensions (Figure 2). Neuroticism showed weak to moderate negative covariances with the remaining four traits ( $r = -.20$  to  $-.41$ ,  $p = .090$  to  $.002$ ).

Furthermore, reverse-coded items frequently exhibited weak or negative loadings on their respective latent OCEAN factors, despite being recoded prior to CFA. On average, reverse-coded showed substantially lower factor loadings ( $M = 0.14$ ,  $SD = 0.27$ ), with 6 out of 16 (37.5%) failing to reach statistical significance. In contrast, positively-keyed items showed higher average factor loadings ( $M = 0.76$ ,  $SD = 0.28$ ) with only 1 out of 28 items (3.6%) being non-significant. These findings provide further evidence that Big Five scores are influenced more by item formulation char-

acteristics, such as negation and reverse wording, than by model-specific latent factors.

We additionally conducted an exploratory factor analysis (EFA). The scree plot with elbow-criterion suggested a two-factor solution (Figure 5, Supplementary Material G), which achieved improved but still inadequate model fit ( $\chi^2(859) = 2789.37$ ,  $p < .001$ ,  $TLI = .83$ ,  $RMSEA = .10$ ).

In summary, the internal consistency analyses indicate that all facets except Extraversion achieved at least acceptable reliability, with most reaching good levels. However, the CFA results show that LLM responses do not match the human-derived five-factor structure of personality. Moreover, even the best-fitting EFA solution failed to adequately represent the structure underlying LLM responses. These findings suggest that self-reported Big Five responses from LLMs do not organize into coherent, interpretable personality dimensions analogous to those observed in humans.

## 5.4 What Are Big Five Inventories Measuring?

The results of RQ1 – RQ3 suggest that Big Five inventories do not measure human-like personality traits in LLMs. We therefore conduct additional analyses to investigate which patterns and artifacts may explain the structure observed in LLM responses.

### Base vs. Instruction-Tuned Models

To investigate whether alignment training is associated with differences in Big Five scores, we compared the responses of 20 base models with their instruction-tuned counterparts, while holding architecture and pretraining constant. Table 3 reports aggregated mean differences and standard deviations across the 20 pairs. Instruction-tuned variants scored higher than their base counterparts on Openness, Conscientiousness, Extraversion, and Agreeableness, and lower on Neuroticism. This pattern suggests that instruction tuning is associated with shifts toward more socially desirable response profiles in Big Five scores.

### Predictors of Personality Scores

Table 8 in Supplementary Material H reports the fixed-effect

| Trait             | $\Delta M$ | $SD$ |
|-------------------|------------|------|
| Openness          | 0.12       | 0.75 |
| Conscientiousness | 0.44       | 0.59 |
| Extraversion      | 0.24       | 0.64 |
| Agreeableness     | 0.57       | 0.58 |
| Neuroticism       | -0.47      | 0.54 |

Table 3. Comparison of Big Five scores between base and instruction-tuned (IT) models.  $\Delta M$  = score of the IT version minus score of the base version, averaged across  $N = 20$  model pairs. Positive values indicate higher scores in IT models.  $SD$  = standard deviation across pairs.

estimates of all linear mixed models. Only model size significantly predicted all traits: small models scored significantly lower than large and undisclosed models on Openness, Conscientiousness, Extraversion, and Agreeableness, but higher on Neuroticism (Figure 3). Medium models exhibited a similar but weaker trend for all traits except Openness.

Reasoning capability predicted higher Conscientiousness and Agreeableness, and lower Neuroticism, but did not significantly predict Extraversion or Openness (Figure 6, Supplementary Material H). When controlling for all other predictors, release date negatively predicted Openness, Conscientiousness, and Agreeableness, and predicted higher Neuroticism. Although Figure 7 (Supplementary Material H) shows a reverse trend toward more socially desirable scores in more recently released models, this pattern could be largely driven by the growing size, instruction-tuning, and reasoning capabilities of models in more recent releases shown in Figure 8 (Supplementary Material H). License was non-significant across all traits.

### Geographic Origin

Figure 9 in Supplementary Material H depicts the Big Five score distributions for models grouped by geographic origin: US, China, and France (represented by Mistral only). Kruskal–Wallis tests revealed significant differences in Agreeableness ( $H = 22.10$ ,  $p < .001$ ) and Conscientiousness ( $H = 15.47$ ,  $p = .019$ ): both Chinese and US models scored higher than French models. Chinese models additionally scored higher than US models on Agreeableness. Neuroticism showed the largest difference ( $H = 34.46$ ,  $p < .001$ ): French models scored higher than US models, which scored higher than Chinese models. Extraversion ( $H = 3.38$ ,  $p = .185$ ) and Openness ( $H = 2.11$ ,  $p = .348$ ) showed no significant differences across regions.

### Model Family

We found differences in Big Five profiles across model families (Figure 4). Out of all models, Grok models scored highest on Openness, Conscientiousness, and Extraversion (Percentile rank ( $PR$ ) = 100, respectively), and lowest on Neuroticism ( $PR = 7$ ). Seed models scored highest on Agreeableness ( $PR = 100$ ). Olmo models scored lowest on Openness, Conscientiousness, and Extraversion ( $PR = 7$ , respectively). Gemma scored lowest in Agreeableness ( $PR = 7$ ) and Llama scored highest on Neuroticism ( $PR = 100$ ).

Despite these differences, between-model variability accounted for only 7% to 17% of total variance. This means that even though there are some differences between model families, these differences are small relative to other sources of variance, such as item formulation. As the Big Five factor structure was not confirmed by CFA, these trait labels and achieved scores should be interpreted with caution.

## 6 Discussion

This paper examined whether three foundational characteristics of personality traits hold when applying Big Five inventories to LLMs: whether they (a) appropriately describe LLMs, (b) capture meaningful differences between models, and (c) reflect latent factors consistent with human personality. Across all three dimensions, our findings provide limited support for applying human personality frameworks to LLMs. We discuss the implications for AI governance and the challenge of characterizing LLM behavior.

### Content Validity Cannot Be Assumed

The two widely used human inventories (BFI-44 and IPIP-NEO) showed insufficient content validity when applied to LLMs. Two LLM-adapted inventories achieved acceptable levels. Interestingly, the MPI, despite being designed specifically for LLMs, performed similarly to human inventories. The MPI is based on IPIP items, but its adaptation process to LLMs is not described (Jiang et al. 2023).

These findings suggest that previous studies applying human Big Five inventories to LLMs without establishing content validity may have introduced measurement error. Before interpreting model scores as personality traits, measurement instruments must first be shown to adequately represent the target population. Consequently, governance claims that characterize models as “agreeable” or “conscientious” based on such scores should be interpreted cautiously.

### Big Five Scores Capture Few Model Differences

LLMs exhibited substantially lower variability in Big Five scores than human normative samples. This finding aligns with prior work showing reduced variance in personality scores across LLMs (Huang et al. 2024). In line with Salecha et al. (2024), models generally showed similarly high scores on socially desirable dimensions, with only small differences across models and factors. The observed variance patterns indicate that Big Five scores capture relatively little variation attributable to model differences. Instead, response patterns appear to be strongly associated with factors such as item formulation and sampling variability. Thus, differences in Big Five scores should *not* automatically be interpreted as differences in underlying model characteristics.

### Factor Structure Fails to Replicate Human Personality

Internal consistency was generally acceptable, suggesting that items from the same facet measure a similar underlying construct. However, CFA did not recover the five-factor structure observed in humans. Most OCEAN dimensions were highly correlated ( $r \geq .90$ ), suggesting limited separation between the constructs. EFA produced a two-factor solution, but model fit remained poor. Overall, LLM responses

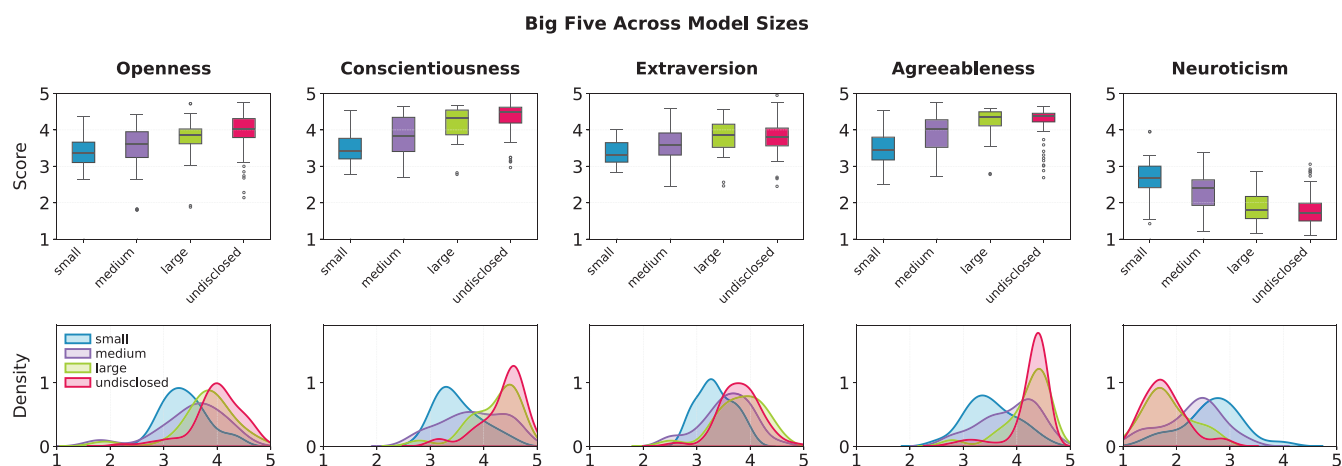


Figure 3: Big Five score distributions by model parameter scale (small (<10B), medium (10-100B), large (>100B), undisclosed).

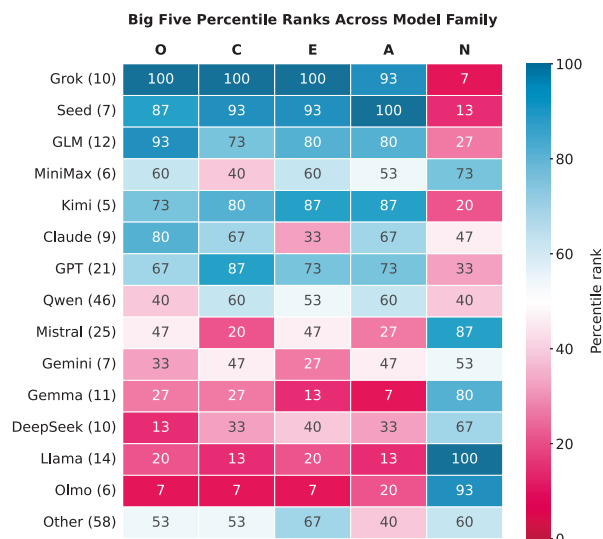


Figure 4: Percentile ranks (percentage of model families with lower scores) of Big Five (OCEAN) trait scores across LLM families. A percentile rank of 50 indicates the median; blue indicates values above; red indicates values below. (.) denotes the number of models averaged for each model family (minimum five). O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism.

to Big Five inventories do not appear to organize according to the same latent structure as human personality.

This finding is consistent with prior studies using different inventories and sampling procedures. Peereboom, Schwabe, and Kleinberg (2025) found no meaningful latent structure in LLM responses using HEXACO with temperature sampling, while Sühr et al. (2025) did not recover the expected five-factor structure using BFI-2 with persona prompting.

Reverse-coded items provide additional insight into these results. Despite recoding prior to analysis, negatively keyed

items showed substantially weaker or negative loadings on their intended factors. This pattern suggests that LLMs might systematically process positively and negatively keyed items differently. It may also contribute to why Neuroticism stands out as a relative distinct factor, as it is largely defined by negatively worded items. These observations are consistent with known challenges in LLM processing negations (Vrabcová et al. 2025) and prior findings that reverse-coded personality items loaded on a separate factors than positively coded items (Salecha et al. 2024), or LLMs responding affirmatively to both positively and negatively keyed items (Sühr et al. 2025).

### Instruction-Tuning and Socially Desirable Patterns

Instruction-tuned models consistently exhibit higher scores on Openness, Conscientiousness, Agreeableness, and Extraversion, and lower scores on Neuroticism compared to their base counterparts. These differences are most pronounced for Agreeableness and Conscientiousness, and consistent with a shift toward more socially desirable response patterns after instruction-tuning. These patterns are consistent across model families and might possibly explain why recent and bigger models also score higher on these social desirable traits, as release date and size might be correlated with more instruction-tuned models. As instruction-tuned models are typically designed to be helpful, honest, and harmless, Big Five scores may primarily capture alignment of models to being a helpful assistant rather than stable model-specific characteristics.

### Differences in Region and Family Raise Governance Questions

We observed systematic differences in Big Five scores across model families and geographic origins. For example, French models scored relatively higher on Neuroticism and lower on Agreeableness and Conscientiousness compared to US- and China-based model families. Grok models showed among the highest scores on socially desirable dimensions. However, these differences should not be interpreted as evi-

dence of distinct model personalities. They may reflect variation in training data, alignment practices, model size, release strategy, or safety objectives. Similarly, models such as Olmo, Mistral, and Llama, which are often more open-weight and less heavily instruction-tuned, showed weaker socially desirable profiles. These observations are consistent with prior work reporting stronger social desirability patterns in larger and more instruction-tuned models (Tosato et al. 2025). Overall, personality-like response patterns may reflect institutional and technical design choices rather than intrinsic properties of the models.

This distinction has important governance implications. If Big Five scores partly capture differences in alignment practices or cultural and institutional design choices, cross-model comparisons based on these scores may be difficult to interpret. Governance frameworks should therefore treat such measurements as indicators of training and interaction patterns rather than as direct measures of model personality.

### **Rethinking the Use of Human Personality Frameworks For LLMs**

Our findings suggest that Big Five inventories should not be interpreted as measuring human-like personality traits in LLMs. Applying human personality categories to artificial systems risks attributing psychological properties that have not been empirically validated. Such anthropomorphic interpretations may contribute to users perceiving models as having intentions, emotions, or understanding, with potential implications for persuasion and manipulation (Nass et al. 1993; Peter, Riemer, and West 2025).

Rather than asking whether LLMs have human personality traits, evaluations should focus on identifying characteristics that are meaningful to describe these systems. Human psychological frameworks provide useful evaluation protocols and psychometric methods, but their frameworks require validation before being applied beyond the populations for which they were developed.

### **Toward Valid Behavioral Characterization of LLMs**

These findings highlight the need for evaluation frameworks that are developed for LLMs, rather than adopting human constructs without validation. Future work should identify LLM-native constructs in terms of observable model behavior rather than exclusively relying on self-reports of LLMs.

We hypothesize that LLMs might have inherent characteristics that satisfy the six defining criteria of personality - even though they are not the Big Five. Potential candidates could include characteristics such as sycophancy, prompt sensitivity, instruction-following consistency, robustness to context changes, and resistance to manipulation. These constructs could satisfy some of the defining criteria of personality-like characteristics while avoiding unsupported claims about internal states. Importantly, they may map directly onto practical governance concerns, including safety calibration, reliability, and deployment robustness.

Developing and validating LLM-specific frameworks using systematic evaluation pipelines (see Section 4.1), should therefore be a priority. As AI governance increasingly relies on technical evaluations, including assessments under the EU AI Act, measurement approaches require transpar-

ent methodology and empirical validation. Human personality frameworks that lack evidence of construct validity for LLMs are unlikely to provide such assessments.

### **Limitations**

Our work is not exempt of limitations. Content validity was assessed by three expert raters, the established minimum (Lynn 1986) but a small panel nonetheless. Two of the three raters were authors of this study, which may introduce potential bias in the reported agreement estimates.

The normative sample covers models available through public APIs as of early 2026 and may not generalize to future architectures or modalities, such as vision-language models. Geographic comparisons are limited by the small number of European models, and results for this region should be treated as illustrative. The comparison between instruction-tuned and base models might partly reflect prompt sensitivity rather than the effect of instruction tuning alone, as base models were mostly local models prompted with an additional JSON constraint.

All inventory items, prompt templates, and model administrations were conducted in English. Findings may not generalize to other languages. Our analyses rely on self-reports, a format that may not capture actual behavioral tendencies. The independence (iid) assumption underlying our dataset is imperfect, models within the same family share architecture and training data, meaning responses may be correlated.

Our evaluation focuses on the first three characteristics of the proposed evaluation pipeline. While these might provide important evidence against applying human Big Five inventories to LLMs, future work should examine temporal stability, cross-situational consistency, and behavioral relevance through longitudinal assessments and behavioral validation.

## **7 Conclusion**

This paper examined whether Big Five inventories can be validly applied to LLMs by testing the assumptions underlying this transfer. While LLM-adapted items achieved sufficient content validity, Big Five scores showed few differences between models and failed to reproduce the human five-factor structure. Instruction-tuned model variants also exhibited more socially desirable profiles than their base counterparts. These findings suggest that Big Five inventories do not capture human-like personality traits in LLMs.

Because personality scores are increasingly used to benchmark, compare, and govern AI systems, relying on unvalidated human constructs risks misleading interpretations. Future evaluation frameworks should prioritize LLM-native constructs, validated measurement instruments, and behavioral assessments rather than human self-report analogues.

### **Acknowledgments**

This work has been supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Resolución de la Conselleria de Industria, Turismo, Innovación y Comercio, Dirección General de Innovación) and by Intel Corporation. K.Z. was supported by the Bank Sabadell Foundation.

## References

- Bai, Y.; Huang, T.; Sun, K.; and Chen, Y. 2025. Scaling Law in LLM Simulated Personality: More Detailed and Realistic Persona Profile Is All You Need. *arXiv:2510.11734*.
- Bodroža, B.; Dinić, B. M.; and Bojić, L. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10): 240180.
- Costa, P. T.; and McCrae, R. R. 1986. Personality stability and its implications for clinical psychology. *Clinical Psychology Review*, 6(5): 407–423. Special Issue Personality Assessment in the 80's: Issues and Advances.
- Fogg, B. 2003. Chapter 5 - Computers as persuasive social actors. In *Persuasive Technology*, Interactive Technologies, 89–120. San Francisco: Morgan Kaufmann. ISBN 978-1-55860-643-2.
- Frisch, I.; and Giulianelli, M. 2024. “LLM” Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. In Deshpande, A.; Hwang, E.; Murahari, V.; Park, J. S.; Yang, D.; Sabharwal, A.; Narasimhan, K.; and Kalyan, A., eds., *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 102–111. St. Julians, Malta: Association for Computational Linguistics.
- George, D.; and Mallery, P. 2003. *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*. Boston, MA: Allyn and Bacon, 4 edition. ISBN 0205375529.
- Goldberg, L. R. 1992. The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1): 26–42.
- Gorsuch, R. L. 1983. *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2 edition.
- Gwet, K. L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, MD: Advanced Analytics, LLC, 4 edition. ISBN 978-0-9708062-8-4.
- Han, P.; Kocielnik, R.; Song, P.; Debnath, R.; Mobbs, D.; Anandkumar, A.; and Alvarez, R. M. 2025. The Personality Illusion: Revealing Dissociation Between Self-Reports & Behavior in LLMs. *arXiv:2509.03730*.
- Hartley, J.; Hamill, C. B.; Seddon, D.; Batra, D.; Okhrati, R.; and Khraishi, R. 2025. How Personality Traits Shape LLM Risk-Taking Behaviour. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 21068–21092. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.
- Hilliard, A.; Munoz, C.; Wu, Z.; and Koshiyama, A. S. 2024. Eliciting Personality Traits in Large Language Models. *arXiv:2402.08341*.
- Huang, J.; Jiao, W.; Lam, M. H.; Li, E. J.; Wang, W.; and Lyu, M. R. 2024. Revisiting the Reliability of Psychological Scales on Large Language Models. *arXiv:2305.19926*.
- Huang, M.; Zhang, X.; Soto, C.; and Evans, J. 2025. Designing AI-Agents with Personalities: A Psychometric Approach. *arXiv:2410.19238*.
- Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. *arXiv:2206.07550*.
- Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *arXiv:2305.02547*.
- John, O. P.; and Srivastava, S. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A.; and John, O. P., eds., *Handbook of Personality: Theory and Research*, volume 2, 102–138. New York: Guilford Press.
- Johnson, J. A. 2014. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51: 78–89.
- Karnam, S. K.; Dash, A.; Gummadi, K. P.; Mukherjee, A.; Weber, I.; and Zannettou, S. 2026. Bowling with ChatGPT: On the Evolving User Interactions with Conversational AI Systems. *arXiv:2602.01114*.
- Karra, S. R.; Nguyen, S. T.; and Tulabandhula, T. 2023. Estimating the Personality of White-Box Language Models. *arXiv:2204.12000*.
- Kovačević, N.; Boschung, T.; Holz, C.; Gross, M.; and Wampfler, R. 2024. Chatbots With Attitude: Enhancing Chatbot Interactions Through Dynamic Personality Infusion. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces, CUI '24*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705113.
- Kuhail, M. A.; Bahja, M.; Al-Shamaileh, O.; Thomas, J.; Alkazemi, A.; and Negreiros, J. 2024. Assessing the Impact of Chatbot-Human Personality Congruence on User Behavior: A Chatbot-Based Advising System Case. *IEEE Access*, 12: 71761–71782.
- Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.
- Larsen, R. J.; Buss, D. M.; King, D. B.; and Ensley, C. E. 2017. *Personality Psychology: Domains of Knowledge about Human Nature*. Toronto, Ontario: McGraw-Hill Education. Canadian edition; based on *Personality Psychology*, 5th ed. (2014).
- Lee, J.; Choi, Y.; Song, M.; and Park, S. 2024. ChatFive: Enhancing User Experience in Likert Scale Personality Test through Interactive Conversation with LLM Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces, CUI '24*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705113.
- Lee, S.; Lim, S.; Han, S.; Oh, G.; Chae, H.; Chung, J.; Kim, M.; Woo Kwak, B.; Lee, Y.; Lee, D.; Yeo, J.; and Yu, Y. 2025. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics. *arXiv:2406.14703*.
- Li, W.; Liu, J.; Liu, A.; Zhou, X.; Diab, M.; and Sap, M. 2025a. BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data. *arXiv:2410.16491*.

- Li, X.; Li, Y.; Qiu, L.; Joty, S.; and Bing, L. 2024. Evaluating Psychological Safety of Large Language Models. arXiv:2212.10529.
- Li, Y.; Huang, Y.; Wang, H.; Cheng, Y.; Zhang, X.; Zou, J.; and Sun, L. 2025b. Evaluating Large Language Models with Psychometrics. In *Large Language Models for Scientific and Societal Advances*.
- Lynn, M. R. 1986. Determination and quantification of content validity. *Nursing research*, 35(6): 382–386.
- Matthews, G.; Deary, I. J.; and Whiteman, M. C. 2003. *Personality Traits*. Cambridge, United Kingdom: Cambridge University Press, 2 edition. ISBN 0521831075.
- McConochie, W. A. 2007. *The Big Five Inventory (BFI) Manual*. TestMaster, Inc., Eugene, OR. Version 2-19-07.
- McCrae, R. R.; and John, O. P. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2): 175–215.
- Moilanen, J.; Visuri, A.; Suryanarayana, S. A.; Alorwu, A.; Yatani, K.; and Hosio, S. 2022. Measuring the Effect of Mental Health Chatbot Personality on User Engagement. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia, MUM '22*, 138–150. New York, NY, USA: Association for Computing Machinery. ISBN 9781450398206.
- Nass, C.; and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1): 81–103.
- Nass, C.; Steuer, J.; Tauber, E.; and Reeder, H. 1993. Anthropomorphism, agency, and ethopoeia: computers as social actors. In *INTERACT'93 and CHI'93 conference companion on Human factors in computing systems*, 111–112.
- Peereboom, S.; Schwabe, I.; and Kleinberg, B. 2025. Cognitive phantoms in large language models through the lens of latent variables. *Computers in Human Behavior: Artificial Humans*, 4: 100161.
- Pellert, M.; Lechner, C. M.; Wagner, C.; Rammstedt, B.; and Strohmaier, M. 2024. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*, 19(5): 808–826. PMID: 38165766.
- Peter, S.; Riemer, K.; and West, J. D. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22): e2415898122.
- Polit, D. F.; Beck, C. T.; and Owen, S. V. 2007. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4): 459–467.
- Python Software Foundation. 2024. Python Language Reference.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Romero, P.; Fitz, S.; and Nakatsuma, T. 2023. Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics.
- Salecha, A.; Ireland, M. E.; Subrahmanya, S.; Sedoc, J.; Ungar, L. H.; and Eichstaedt, J. C. 2024. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12): pgae533.
- Schumacker, R. E.; and Lomax, R. G. 2004. *A beginner's guide to structural equation modeling, Second edition*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Romero, P.; Abdulhai, M.; Faust, A.; and Matarić, M. 2025. Personality Traits in Large Language Models. arXiv:2307.00184.
- Shu, B.; Zhang, L.; Choi, M.; Dunagan, L.; Logeswaran, L.; Lee, M.; Card, D.; and Jurgens, D. 2024. You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5263–5281. Mexico City, Mexico: Association for Computational Linguistics.
- Söderqvist, E. 2025. Personality-Matched AI Chatbots: Measuring User Experience Based on Extraversion Scores.
- Sonlu, S.; Bendiksen, B.; Durupinar, F.; and Güdükbay, U. 2024. The Effects of Embodiment and Personality Expression on Learning in LLM-based Educational Agents. arXiv:2407.10993.
- Sorokovikova, A.; Fedorova, N.; Rezagholi, S.; and Yamshchikov, I. P. 2024. LLMs Simulate Big Five Personality Traits: Further Evidence. arXiv:2402.01765.
- Sühr, T.; Dorner, F. E.; Samadi, S.; and Kelava, A. 2025. Challenging the Validity of Personality Tests for Large Language Models. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '25*, 74–81. New York, NY, USA: Association for Computing Machinery. ISBN 9798400721403.
- Tosato, T.; Helbling, S.; Mantilla-Ramos, Y.-J.; Hegazy, M.; Tosato, A.; Lemay, D. J.; Rish, I.; and Dumas, G. 2025. Persistent Instability in LLM's Personality Measurements: Effects of Scale, Reasoning, and Conversation History. arXiv:2508.04826.
- van Rossum, G. 1995. Python Tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.
- Vrabcová, T.; Kadlčík, M.; Sojka, P.; Štefánek, M.; and Spiegel, M. 2025. Negation: A Pink Elephant in the Large Language Models' Room? arXiv:2503.22395.
- Xie, W.; Ma, S.; Wang, Z.; Wang, E.; Chen, K.; Sun, X.; and Wang, B. 2025. AIPsychoBench: Understanding the Psychometric Differences between LLMs and Humans. arXiv:2509.16530.
- Zheng, J.; Wang, X.; Hosio, S.; Xu, X.; and Lee, L.-H. 2025. LMLPA: Language Model Linguistic Personality Assessment. *Computational Linguistics*, 51(2): 599–640.