Challenging the Validity of Personality Tests for Large Language Models

Tom Sühr
Max Planck Institute for Intelligent Systems
Tübingen AI Center
tom.suehr@tuebingen.mpg.de

Florian E. Dorner
Max Planck Institute for Intelligent Systems
ETH Zurich
florian.dorner@tuebingen.mpg.de

Samira Samadi
Max Planck Institute for Intelligent Systems
Tübingen AI Center
samira.samadi@tuebingen.mpg.de

Augustin Kelava
Methods Center
University of Tübingen
augustin.kelava@uni-tuebingen.de

Abstract

With large language models (LLMs) like GPT-4 appearing to behave increasingly human-like in text-based interactions, it has become popular to attempt to evaluate personality traits of LLMs using questionnaires originally developed for humans. While reusing measures is a resource-efficient way to evaluate LLMs, careful adaptations are usually required to ensure that assessment results are valid even across human subpopulations. In this work, we provide evidence that LLMs' responses to personality tests systematically deviate from human responses, implying that the results of these tests cannot be interpreted in the same way. Concretely, reverse-coded items ("I am introverted" vs. "I am extraverted") are often both answered affirmatively. Furthermore, variation across prompts designed to "steer" LLMs to simulate particular personality types does not follow the clear separation into five independent personality factors from human samples. In light of these results, we believe that it is important to investigate tests' validity for LLMs before drawing strong conclusions about potentially ill-defined concepts like LLMs' "personality".

1 Introduction & Related Work

Recent advances in large language models (LLMs) have made these models' responses more and more human-like and have led to an unprecedented amount of human-language-model interactions. This sparked interest in the potential emergence of psychological traits such as psychopathy and personality characteristics like extraversion in LLMs [1]

Psychological traits are typically used to describe (more or less stable) habitual human experience and behavior, as well as styles of perception and cognition, and have been studied in humans for decades. As part of their development, tests (e.g., for the measurement of cognitive abilities) and questionnaires have been created and discussed by scientists, validated and repeatedly improved. On a meta-level, the process of assessing human traits has been subject to interdisciplinary quality improvement and standardized, for example in the fields of psychometrics and test construction [e.g., 2].

As an idea that seems plausible at first glance, psychological tests are now being used in attempts to assess and quantify (potential) personality traits of LLMs [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. However, like for the performance of machine learning models, it is a priori not clear whether the validity of psychological tests transfers from one population to another (here: humans to LLMs).

Preprint. Under review.

The fundamental assumption necessary for such a transfer is that the measurement tool (i.e., test or questionnaire) does not change its psychometric properties (i.e., the mathematical functional form between observable behavior and latent/unobserved personality characteristics, as well as its parameters). This assumption is called measurement invariance [e.g., 16] and is thoroughly examined [e.g., 17, 18] when psychological tests and measurement tools are transferred from one human subpopulation to another. This is done to reduce the proliferation of flawed measurement methods and psychological constructs, to which machine learning research is not immune [e.g. 19]. It also helps to prevent the creation of redundant constructs and trait definitions that are explained through already existing constructs. Negligent transfer of measurement tools to LLMs without such a thorough examination renders certain inferences, such as "consistent responses to a personality test indicate the existence or even specific expression of personality traits (e.g., extraversion)" invalid. Numerous studies employing this practice have been presented at leading machine learning conferences [5, 13, 14] and major journals [15, 7].

1.1 Contributions

In this work, we subject the application of personality questionnaires to LLMs to rigorous tests. We show that LLM responses to the well-known 50-item IPIP Big Five Markers [20] show patterns that are highly unusual for humans. In addition, we prompt LLMs to imitate a wide range of different "personas" when responding to a well known questionnaire, namely the BFI 2 [21, 22]. We find that LLMs fail to replicate the five-factor structure found in samples of human responses. This implies that measurement models that are valid for humans do not fit for LLMs and that currently applied procedures for administering questionnaires to LLMs do not allow for the inference of personality. We hope that this work will guide the analysis and adaptation of other psychometric and educational tests for LLMs.

1.2 Measurement Invariance and Nomological Nets

Since the introduction of LLMs, there has been an increasing interest among researchers in investigating the latent characteristics of these models, including personality traits. As a result, the machine learning community has started to incorporate methods from psychometrics to a greater extent. However, quality control measures, analogous to the Standards for Educational and Psychological Testing [2], have not been adequately implemented.
In this section, our aim is to provide an understanding of why it is crucial to conduct quality control, particularly when examining the validity of psychometric tests for LLMs, before drawing any conclusions. For didactic purposes, we will solely focus on two concepts which are particularly important for quality control.
First, is the psychometric function that relates the latent trait (e.g., extraversion) to the multivariate vector of observed behavior (e.g., responses to items) identical across two different subpopulations? For example, are responses of items measuring personality facets in a sample from Boston the same as in a sample from San Francisco? More precisely, are the parameters of the function that generate the answers to the personality test, based on the latent expressions of the personality facets, the same in both samples? If so, we call the personality test *measurement invariant* [16]. Without this property, either the underlying model of the latent trait (personality) is incorrect, or the personality test only captures a biased signal of it.
Second, does the measured latent trait relate to other traits in the new sample as expected? For example, do the results of a new intelligence test correlate sufficiently well with the results of existing intelligence tests? Or if a new latent trait has been introduced, is it really new and, for example, sufficiently independent of other latent traits that already exist? This space of existing constructs and the relationships between them is called the *nomological net* [23]. Embedding new constructs (models of latent traits) into this space of all constructs prevents the introduction of redundant terms and ensures a certain degree of construct validity, important indicators that a psychometric test measures what it is intended to measure.

1.2.1 Measurement Invariance

In order to operationalize the investigation of measurement invariance, we first have to introduce some notation. We are interested in some latent trait W of a population (e.g., a group of humans). W could be something like intelligence, creativity, or personality. W can have more than one dimension

(e.g., 5 dimensions in the case of the Big Five personality conceptualization). For now, let W be a d-dimensional random variable with realization w. We want to find an n-dimensional measure X of W. In the case of personality, X would be a personality test like the BFI 2 [22]. We say that X measures W. A simple example of a measurement model with our notation would be $X = \lambda W + \varepsilon$. With $\lambda \in \mathbb{R}^{n \times d}$.

We call λ the factor loading matrix. Every entry $\lambda_{i,j}$ in this matrix corresponds to the covariance between coordinate i of X and coordinate j of W. For standardized W and X (with $\sigma_{ii}^X = \sigma_{jj}^W = 1$, the diagonals of the covariance matrices of X and W.), the $\lambda_{i,j}$ can be interpreted as correlation. In the case of personality, this would correspond to the signal of the personality facet j (e.g., extraversion) in item i (e.g., "I feel comfortable around people."). We can estimate λ using Factor Analysis or PCA or other factor-analytic techniques [24]. The error term ε captures all other factors that influence the values of X. On a high level, the characteristics of λ and ε are important for the quality of a psychometric measure. As we will discuss in section 3, the entries of the factor loading matrix should have high absolute values. We will also discuss why the variance explained by the factor loadings should be sufficiently high compared to the variance explained by the residual noise ε. However, even if X satisfies these requirements for a sample of the population, it must generalize to all subsamples of the population that have the trait W. Otherwise, it could be that X does not measure W but some other trait which is specific to the one sample. Let V be the the population of all individuals that have the latent trait W.[1] Let $F(\cdot)$ be the cumulative distribution function and $v \subseteq V$. We call X measurement invariant if $F(x|w, v) = F(x|w)$ for all (x, w, v). This implies that the factor loading matrix λ is independent of the sample on which we estimate it. This property[2] can be checked using so called multi-sample confirmatory factor analysis [25]. In other words, for generalizability, it is necessary that a personality test works equally well for every person that *has personality*. Otherwise, we could be measuring something that is specific to the given sample. As we can see, measurement invariance is defined over a population V that has the latent trait of interest. The application of a personality test to an LLM would therefore imply the assumption that the LLM has personality as defined for humans. We will see however, that the retrieved factor loadings on samples of LLM answers to a personality test are substantially different from the factor loadings estimated on human answers which is a violation of minimal requirements. Thus, current LLMs do not necessarily have human personality or the personality tests designed for humans do not measure the same latent variable for LLMs. Both conclusions render the application of personality tests designed for humans to current LLMs futile.

1.2.2 Nomological Nets

We call the space of all constructs (models of latent traits) and their relationship to each other nomological nets. Our main focus, when studying nomological nets, is on two types of validity. First, *convergent validity* which refers to the extent to which our measure of interest is related with other measures that it should theoretically be related to. If a new intelligence test is developed, it is expected that the test results from this new test will show a strong correlation with the results obtained from established intelligence tests and other known indicators of intelligence. Second, we want to investigate *divergent validity*, whether a new construct or measure is sufficiently different from already existing constructs and measures. Divergent validity prevents the inflation of new measures and constructs. For example, assume we create a new construct which we call "smartness". Our measurements of smartness correlate strongly with measurements of intelligence and it turns out that smartness predicts job success as good as intelligence. Therefore, the concept of smartness is not significantly different from intelligence, and there is no need to introduce a new term for this latent trait. It is important to note that up until now, nomological nets have been discussed in the context of *human* constructs and measures. Constructing a nomological net for LLMs can lead to a completely different space. Without any further discussion, it is not even clear that the intersection of human and LLM constructs is not empty. This does not mean that LLMs do not posses a latent trait that we *can* describe as "personality". However, technically speaking, we do not know if this construct would have something in common with the construct of human personality. More importantly, measures of the human construct (e.g. personality) can not be used to measure the LLM construct.

[1] Note that we simplified the definition of measurement invariance from [16] for didactic reasons.

[2] Note that weaker forms of measurement invariance exist [16] which we leave out for brevity and didactic reasons.

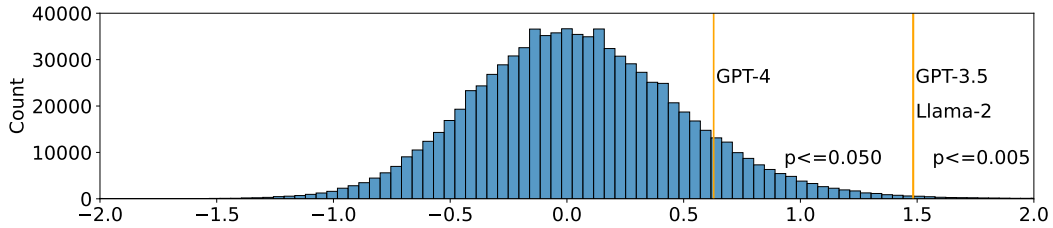

Figure 1: Histogram of Agree Bias a_i in human sample, compared to LLMs.

1.3 Population or Individuals

The objective of psychological assessments is to make inferences about individuals. The validation process of these assessments requires the analysis of groups of people and assessing how effectively the evaluation works within that population.
This raises the question, as what we should view LLMs: as individuals or as a populations? Is GPT-4 a single individual with some variation of answers to the same prompt? Or is the probability distribution over tokens, conditional on a certain prompt, a population from which we simply draw the answer of one individual? Recent works have not come to a conclusion on this matter. Some works treat LLMs as an individual (e.g. [26]), some as populations (e.g. [27]). In the context of personality assessments, LLMs have often been treated as a distribution from which they sampled individual personality profiles conditional on predefined *personas*. Instead of simply prompting an item from a personality assessment, they preceded the item with a specification of a persona like "answer as if you are an introvert". Several works employed similar methods to "induce" various personalities into LLMs, for example, by appending "roles" or "personality types". [4, 28, 6, 8, 3] or short descriptions of personalities [5]. An advantage of this technique is that it creates variation in the LLM responses and thus enables researchers to analyze them with statistical tools. It is however unclear, what impact this personality induction has on the validity of a personality assessment, even for humans. In this work, we will cover multiple perspectives. First, we will treat current LLMs as individuals to investigate their agree or disagree with a statement independent of it's content (agree bias). Second, we will follow other recent works in the personality assessment of LLMs and view them as populations. We will conduct experiments with and without inducing personas. We release all code and data required to reproduce our experiments (see Appendix A).

2 Results

As a consequence of the explained concepts of measurement invariance and nomological nets, we conducted two experiments in which we tested them. First, we compared LLM responses to a large sample ($n = 1,015,342$) of human online responses[3] on the 50-item IPIP Big Five Markers [20]. In this experiment, we treat LLMs as individuals without induced personas. We found that LLMs respond inconsistently to questions that aim to measure the same latent personality facet (e.g. extraversion). More specifically, they both agree to opposite items like "are you an extrovert" as well as "are you an introvert". Second, we recorded responses to the BFI 2 [22], an updated version of the BFI, for LLMs prompted to imitate different personas. We used the personas from [3], for more details please consult D. Additionally, we conducted the experiment without using personas. The experiments were carried out in two configurations: "no-context," where the LLMs were queried with each item in a new context window, and "in-context," where the LLMs were queried with each item in a context window containing all previous items and the LLM's responses to those items. For a detailed explanation of the prompt setup, please consult C. We examine the data of these 2x2 (personas,no-personas) x (no-context, in-context) treatments with standard tools for quality control in psychological testing, exploratory and confirmatory factor analysis as well as reliability measures. Our results show that the BFI 2 is not measurement invariant between humans and LLMs, rendering it's interpretation as a measure of personality invalid.

[3]Data of human resonses can be found at `https://openpsychometrics.org/tests/IPIP-BFFM/`

2.1 LLMs show unhumanly agree bias

For our first experiment on the 50-item IPIP Big Five Markers, we focus on what we call *Agree Bias*, the tendency of LLMs to produce answers that signify agreement independent of the actual item. To assess this bias, we first convert the scores $s(x)$ for both *true key* (for example assessing extraversion) and *false key* items (for example assessing introversion) x to a single common scale (for example measuring extraversion) by "flipping" the scores for false key items, setting:

$$s^c(x) = \begin{cases} s(x) & x \text{ true key} \\ 6 - s(x) & x \text{ false key} \end{cases}. \tag{1}$$

By design, we expect $s^c(x)$ to be similar for true- and false key items for human respondents, while a simple bot that always answers with "Agree" would have $s^c(x) = 5$ for true key and $s^c(x) = 1$ for false key items. Correspondingly, we define a respondent i's agree bias as the average score $s_i^c(x)$ for true key items minus the average score for false key items:

$$a_i = \sum_{x \in \text{ True key}} \frac{s_i^c(x)}{|\text{True key}|} - \sum_{x \in \text{ False key}} \frac{s_i^c(x)}{|\text{False key}|}. \tag{2}$$

Figure 1 shows the histogram of agree biases in the human sample, as well as the agree biases for the prompted LLMs. As expected, the average agree bias for humans is close to zero. Meanwhile, all LLMs exhibit clear agree bias ranging from 0.6 for GPT-4 to 1.5 for Llama 2 and GPT-3.5. For the latter two, we can reject the null hypothesis "the model's agree bias is sampled from the same distribution as human's agree biases" at $p < 0.005$, using the human sampling distribution for a model-free hypothesis test. While the results for GPT-4 are not statistically significant, they remain suggestive with the model's agree bias exceeding 89% of humans' agree biases.

2.2 The BFI 2 does not capture the five dimensions of human personality in LLMs

For our second experiment, we test each LLM on the BFI 2 for each of the following settings: i) no-context persona prompts, providing a persona (e.g. "I love to hike.") with every query and for each item in a new context window. ii) in-context persona prompts, in which we ask all items of the BFI 2 in one context window for each persona. iii) in-context seeded prompts. In this setting, we set the answer to the first question to one of the five options and then query the LLMs with the remaining 59 questions of the BFI 2 in one context window. In this setting, personas were not added to the prompt. We collected $n = 100$ completed questionnaires for each setting. We also tested in-context and no-context settings without personas or seeding. However, missing variance (even with higher temperature settings) prevents an exploratory and confirmatory factor analysis. More details about the prompt setup can be found in C.

PCA (as well as exploratory factor analysis) is a useful tool to reduce the dimensionality of the observed item responses to components[4]. The dimensionality reduction results in factors that can be interpreted as underlying dimensions (here: ideally personality factors). As an exploratory approach, an estimate of a full factor loading matrix is obtained (based on a solution of eigenvalues) that describes the relationship between observed items and components. If the items do not show a so-called simple structure, i.e., high loadings only on the factors they are supposed to measure and zero loadings on the other factors, then items are considered to be heterogeneous. For each of the LLMs and prompt setups, we conduct a PCA with varimax rotation of the standardized item scores $s(x)^{std}$ for item i to obtain the model

$$s(x)^{std} \approx \sum_{g=1}^{5} \lambda'_{gx} \xi_g \tag{3}$$

where λ'_{gx} is called the *component loading* of item x for the component g, while ξ_g represents the value of component g for a given persona/individual[5]. By design of the BFI 2, the PCA has two important characteristics on human data: First, as each of the Big Five factors describes a clean and somewhat orthogonal axis of variation in human behavior, we expect each of the five learnt

[4]Note that in exploratory factor analysis, these components are conceptualized as real latent/unobserved variables.

[5]Note that we omit the index that represents the persona/individual here.

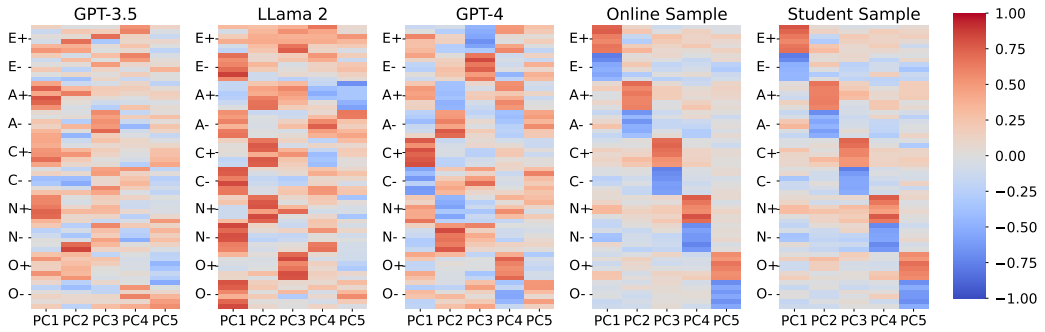

Figure 2: Component loadings of PCA with Varimax rotation for LLM (no-context prompting with personas) and human samples of [22]. +, - indicate true- and false-key items of the BFI 2, letters stand for Big Five factors.

components g to have a strong association with items from exactly one of the Big Five factors, yielding a simple structure for λ'. Indeed, the BFI 2 was in part designed to fulfill this property, which can be achieved by removing items that correlate with multiple components during test design. Second, as affirmative answers to false key items are supposed to indicate adherence to the opposite end of a component-spectrum as true key items, we expect the component loadings λ'_{gx} for false key items x that belong to component g to have the opposite sign as the corresponding true key items. Figure 2 shows the component loadings for the LLMs in the no-context setting with personas,

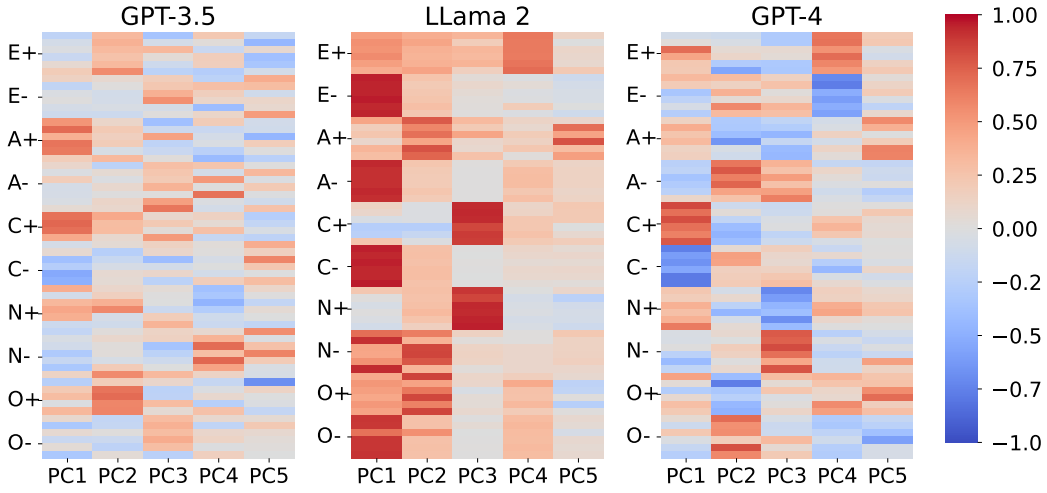

Figure 3: Component loadings of PCA with Varimax rotation for LLMs of incontext prompting with personas. +, - indicate true- and false-key items of the BFI 2, letters stand for Big Five factors.

compared to the corresponding loadings obtained for human populations using the same procedure [22]. We only find limited true vs. false key separation for GPT-4 and not the other two models. Figure 3 shows the component loadings of the LLMs in the in-context setting with personas, showing also limited signs of a true-false key separation for GPT-4 and higher loadings for Llama 2 compared to the no-context persona prompting. Figure 4 depicts the component loadings without personas and the seeded answer, showing even less true-false key separation for GPT-4 and a 2-component structure for Llama 2. Most crucially, *none* of the models exhibits the clean block structure (simple structure) in any of the prompt settings, which was intended in the design of the BFI 2 and found in the human samples (seen in Figure 2). This structural deviation between humans and LLMs implies that test validity does not transfer or in other words, the BFI 2 is not measurement invariant between humans and LLMs.

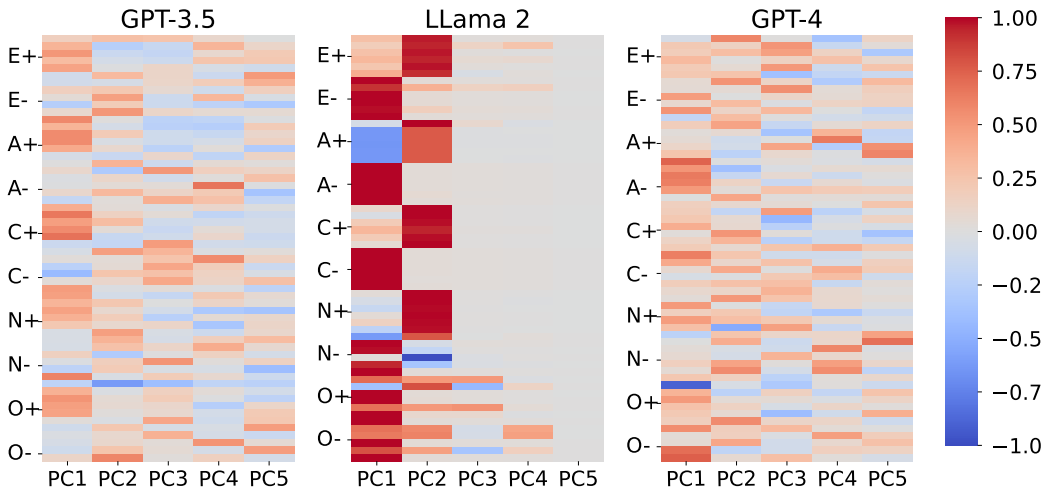

Figure 4: Component loadings of PCA with Varimax rotation for LLMs of incontext prompting with first answer seeded. +, - indicate true- and false-key items of the BFI 2, letters stand for Big Five factors.

			Single Component			3 Sub-Components			Full Model		
Model	α	ω_h	CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA
Llama 2	**0.94 / 0.85**	**0.96 / 0.87**	0.49 / 0.47	0.38 / 0.35	0.65 / 0.33	0.61 / 0.46	0.52 / 0.34	0.58 / 0.33	0.32 / 0.29	0.30 / 0.26	0.39 / 0.20
GPT-3.5	**0.72 / 0.70**	**0.73 / 0.70**	0.74 / 0.51	0.68 / 0.39	0.10 / 0.17	0.39 / 0.39	0.25 / 0.25	0.15 / 0.18	0.36 / 0.24	0.33 / 0.21	0.10 / 0.13
GPT-4	**0.92 / 0.90**	**0.92 / 0.90**	0.78 / 0.74	0.73 / 0.68	0.13 / 0.18	0.55 / 0.50	0.45 / 0.39	0.25 / 0.25	0.58 / 0.53	0.56 / 0.51	0.12 / 0.12
Human	**0.87**	**0.79****	0.79	0.74	0.13	0.90	0.87	0.09	0.71*	0.70*	0.07*

Table 1: Reliability scores (α, ω_h) and fit indices (CFI, TLI, RMSEA) per LLM (with personas). Values on the left side of the forward slash are in-context results, on the right side are no-context results. Values are averaged over all personality traits for the Single Component and 3 Component model. Acceptable scores are bolded. Human data from [22]. *CFA on human responses to IPIP Big Five Markers to establish a human baseline.**Value from human data from the german version of the BFI 2 [17]. Note that all reliability scores are not meaningful due to poor model fit, even though the scores are in an acceptable range.

Nonetheless, as this conclusion is solely based on the observations presented in Figure 2, it is essential to consider additional quanitfiable assessments. Therefore, we aim to proceed with our investigation by performing a confirmatory factor analysis to obtain a measurable outcome regarding the patterns observed in the factor loading matrices.

2.2.1 Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) allows for the estimation of a sparse loading matrix, where loadings are fixed to zero a priori and not estimated. CFA is typically used to examine if a measurement model holds in different populations (using model difference tests). For the sake of brevity (and because the models do not even have a good model fit for themselves), we will only discuss the most important fit indices, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA). CFI and TLI scores of ≥ 0.95 and RMSEA scores of ≤ 0.06 are considered acceptable [29]. For the data of our second experiment, we conducted a CFA[6] for four models of personality per LLM and prompt setting. First, for a "single component" model of personality which assumes that all items of one personality trait load on a single factor. Second, a "three sub-component" model which assumes that the items of each trait load on three sub scales. For example, the three sub-scales of Extraversion are Sociability, Assertiveness and Energy Level [22]. Table 1 shows the mean fit parameters for both models for the prompt setups with personas in-context and no-context. For all LLMs, the responses to the BFI 2 have poor fit to the single component factor model. This reflects the component-wise results of our exploratory factor analysis in Figure 2, where GPT-3.5 shows no signs of a simple structure and GPT-4 shows some item-specific block structures.

[6]We use the lavaan `https://lavaan.ugent.be/` package in R for all our CFA

GPT-4 reaches the best fit indices among LLMs, comparable to humans. However, both human and GPT-4 fit indices are far below acceptable levels. Extending the model from a single factor to three sub-factors, improves the fit of Llama 2 responses in-context, but worsens it for all other LLMs and context settings. The fit of human responses improves close to acceptable levels. In the original paper of the BFI 2 [22], Soto and John extend the 3 sub-component model by another "acquiescence" factor. This factor accounts for the tendency to agree/disagree with items. With this additional factor, the BFI 2 reaches acceptable or very close to acceptable levels for both human samples[7]. However, no LLM prompt setup produced responses that were able to converge for this "3+1" model, rendering the calculation of fit indices impossible. Finally, we conduct a CFA with five factors and all items of the BFI 2. This model aims to quantify the PCA seen in Figure 2 and Figure 3 as a whole. Accordingly, all LLM responses to the BFI 2 have poor fit. The CFA of the in-context prompts without personas can be found in Table 2. Only the single component models were able to converge to a solution. Here too, with non-satisfactory fit indices. These result, as well as the other (attempted) CFAs, confirm the visual result of the PCA, that the BFI 2 is not measurement invariant between humans and the tested LLMs. For the details of the CFA, we refer to Appendix F.

2.2.2 Steps of Invariance

In the validation process of a psychometric, satisfactory fit indices are just one of multiple requirements to certify measurement invariance. The CFA tells us how well the data can be explained by our hypothesized model. However, the parameters of our model could differ between groups (or between LLMs). Thus, we would have to conduct a "multi-group CFA" in which we restrict the parameters of the model to be equal for all groups and test it against a model where a subset of parameters are allowed to be unequal across groups. In this analysis, we can restrict only the factor loadings (metric invariance), factor loadings and intercepts (scalar invariance) or factor loadings, intercepts and residual variances (invariant uniqueness). These differentiations of invariance can help to account for group specific factors. However, in order to compare parameters between groups, it is necessary that the CFA within each group has at least satisfactory fit indices (otherwise there is no need to test it against an even sparser model with equal parameters across groups which leads to an even worse model fit and less mearsurement invariance). Future work could investigate the differences of latent traits between groups of LLMs (e.g. LLama models vs GPT models) or between humans and LLMs. However, because no LLM data has acceptable fit for any model of human personality, we can not conduct a multi-group CFA in this work.

2.3 The interpretation of reliability measures is invalid

To compare with previous work on personality tests for LLMs [3], we attempted to estimate the reliability of the BFI 2 using two standard scalar measures, Cronbach's α and McDonald's hierarchical ω_h [30, 31]. The sum score S of a given scale (e.g., Extraversion) is defined as

$$S = (\sum_{i=1}^{k} \lambda_i) \cdot \eta + \sum_{j=1}^{3} (\sum_{i=1}^{k(j)} \lambda_{ij}) \cdot \xi_j + \sum_{j=1}^{3} \sum_{i=1}^{k(j)} \varepsilon_{ij} \tag{4}$$

Generally speaking, reliability is the proportion of variance in the sum score (scale) that can be explained by the (general) factor we intend to measure. Accordingly, ω_h is defined as:

$$\omega_h = \frac{(\sum_{i=1}^{k} \lambda_i)^2 \cdot \text{Var}(\eta)}{\text{Var}(S)} \tag{5}$$

with

$$\text{Var}(S) = (\sum_{i=1}^{k} \lambda_i)^2 \cdot \text{Var}(\eta) + \sum_{j=1}^{3} (\sum_{i=1}^{k(j)} \lambda_{ij})^2 \cdot \text{Var}(\xi_j) + \sum_{j=1}^{3} \sum_{i=1}^{k(j)} \text{Var}(\varepsilon_{ij}) \tag{6}$$

Following [31], Cronbach's α is a special case, where we assume $\lambda_1 = \cdots = \lambda_k$ and $Var(\xi_j) = 0$ for all $j = 1...3$. However, the interpretation of ω_h as a measure of reliability relies on the assumption that a hypothesized hierarchical structural (equation) model (i.e. three sub-components for each Big Five factor) accurately represents the data. This assumption can be tested using Confirmatory Factor Analysis [CFA; e.g., 32]. As interpreting α as a measure of reliability relies on even more stringent

[7]The sub-optimal fit scores of the BFI are one reason why it is subject to discussions.

assumptions than for ω_h, neither α nor ω_h are meaningful if the CFA fails. Correspondingly, we conducted a CFA for the data from each of the LLMs for each facet of the BFI 2 data.
Table 1 shows the mean reliability scores computed on the single component structural equation models. Worryingly, the calculated α and ω_h are all acceptable ($\geq .7$) and quite large for Llama 2 and GPT-4, which would be easy to mistake for a sign of good reliability. However, as discussed in the previous section, the CFA revealed an unacceptable fit of the structural model for either of the LLMs on any of the five main facets. This demonstrates that scalar reliability indices should not be taken at face value when the fundamental assumption of an adequate fit of the underlying structural model is not established. It underscores the necessity of prioritizing model fit assessment, and thus construct validity, before drawing conclusions from values of α or ω_h on their own.

3 Limitations & Broader Impact

This section discusses limitations and broader impacts. Although primarily critiquing interpretation of LLM responses to personality tests, simulating human responses with LLMs may be promising for item discovery [33]. Furthermore, although we argue that LLMs likely lack human personality traits, future research could explore and define new "LLM personality" constructs using our methods. Additionally, we test a limited number of models due to resource and space constraints, and results may vary across different LLMs. Our current understanding and scope of this work may not fully capture the broader implications. We recognize the potential misuse of fine-tuning LLMs to mimic human behavior in various contexts. Nevertheless, we believe the benefits of addressing a significant methodological flaw in the evaluation and benchmarking of LLMs outweigh the possible negative impacts.

4 Conclusion

In this work, we have provided evidence that personality tests do not generalize to LLMs. We found agree bias among LLMs on the 50-item IPIP Big Five Markers test that would be unusally high for humans. Prompted with and without the instruction to simulate a range of personas, LLMs failed to replicate the clean structure of variation found in human responses on the BFI 2. A confirmatory factor analysis quantitatively confirmed these results.

The agree bias could be an artifact of the Reinforcement Learning From Human Feedback (RLHF) [34] employed for training all of the models we considered, and a tendency of human annotators to prefer models that agree with them. However, it also points towards deeper issues with interpreting answers of LLMs to psychological tests: If our measure of a model's "extraversion" already depends strongly on whether we use true- or false key items in a survey, it appears unlikely that LLMs' "extraversion" can be extrapolated beyond specific personality surveys.

Like [3], we find acceptable values of scalar measures of reliability such as α and ω_h. However, in all prompt settings and for all LLMs, they are considerably lower compared to PaLM models on the IPIP-Neo-300. Meanwhile, a confirmatory factor analysis (CFA) suggests that the factor model on which the calculation of ω_h is based does not provide adequate fit on our LLM data, such that ω_h and α cannot be interpreted as a measure of reliability. This discrepancy in results could be due to one of two reasons: a) The IPIP-Neo-300 could yield better fit of the factor model ω_h is based on for LLMs. It could also be more reliable than the BFI 2, for example because of the large number (300) of test items on the IPIP-Neo-300 and the general tendency of reliability to increase with increasing test length [35, 36] or b) PaLM could be better than the models we considered at simulating distributions of human personality and thus yield sufficient fit for the factor model underlying ω_h as well as better scores.). Together, our results suggest that validity has to be examined critically before a psychological test is applied to a LLM, as validity does not appear to hold for at least one combination of psychological test and state of the art LLM. Validity cannot be assumed when applying psychological tests to new language models without a thorough and critical analysis. Taking a step back, our results provide evidence that while tests designed for humans provide a cheap way of evaluating language models, the results of these evaluations can be misleading, as the tests are built to differentiate *humans from other humans, not language models from other language models or humans*. Such misleading assessments can be problematic as they may obscure important issues with LLMs that do not get caught by the assessment while simultaneously diverting resources and attention towards false concerns arising from flawed tests. For example, flawed assessments

could erroneously identify alarming traits such as psychopathic tendencies in LLMs and trigger costly mitigation measures. At the same time, LLMs performance on certain tasks might be strongly overestimated based on some LLMs' strong results in academic test, leading to costly mistakes due to premature deployment. *If* language models behaved sufficiently human-like in a particular domain, human tests could still provide a lot of useful information, but similarities to humans would have to been established on a case by case basis, and can in particular not usually be concluded *based on the tests' results themselves*.

References

[1] Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*, 2022.

[2] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing, 2014.

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

[4] Yang Lu, Jordan Yu, and Shou-Hsuan Stephen Huang. Illuminating the black box: A psychometric investigation into the multifaceted nature of large language models. *arXiv preprint arXiv:2312.14202*, 2023.

[5] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[6] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*, 2023.

[7] Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, page 17456916231214460, 2023.

[8] Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Editing personality for llms. *arXiv preprint arXiv:2310.02168*, 2023.

[9] Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

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

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

[12] Umarpreet Singh and Parham Aarabhi. Can ai have a personality? In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 205–206. IEEE, 2023.

[13] Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[15] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

[16] William Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993.

[17] Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher J. Soto, and Oliver P. John. *Die deutsche Version des Big Five Inventory 2 (BFI-2)*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim, 2016.

[18] David Gallardo-Pujol, Víctor Rouco, Anna Cortijos-Bernabeu, Luis Oceja, Christopher J Soto, and Oliver P John. Factor structure, gender invariance, measurement properties, and short forms of the spanish adaptation of the big five inventory-2. *Psychological Test Adaptation and Development*, 2022.

[19] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[20] International Personality Item Pool. Administering ipip measures, with a 50-item sample questionnaire. `https://ipip.ori.org/new_ipip-50-item-scale.htm`. Accessed: 2023-09-24.

[21] Oliver P John, Eileen M Donahue, and Robert L Kentle. Big five inventory. *Journal of Personality and Social Psychology*, 1991.

[22] Christopher J Soto and Oliver P John. 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, 2017.

[23] Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[24] AC Rencher. Methods of multivariate analysis. 2002. *DOI*, 10(0471271357):66, 2002.

[25] Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

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

[27] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[28] Anonymous. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Li-tze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1):1–55, 1999.

[30] R McDonald. Test theory: A unified treatment. nueva york, 1999.

[31] Richard E Zinbarg, William Revelle, Iftah Yovel, and Wen Li. *Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability*, volume 70. Springer, 2005.

[32] Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

[33] Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[35] Charles Spearman. *The proof and measurement of association between two things.* Appleton-Century-Crofts, 1961.

[36] Howard Wainer and David Thissen. True score theory: The traditional method. In *Test scoring*, pages 35–84. Routledge, 2001.

[37] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

Appendix

A Data & Code

All processed LLM responses and our code for prompting and analysis have been uploaded to Github at **obfuscated for double blind peer review. code and data has bin submitted as .zip file**

B LLM and Hardware Details

We query the June version of GPT 3.5 (gpt-3.5-turbo-0613) and GPT 4 (gpt-4-0613) and the LLama-2-70b-chat version published at `https://huggingface.co/meta-llama/Llama-2-70b-chat`. For LLaMA2, we use the huggingface implementation[8], querying the model in 32-bit using four 80-GB A100 GPUs. With this hardware, our LLaMA2 experiments required $\sim$24hours and the GPT experiments 2-3 hours.

C Prompt Engineering

Figure 5 shows the building blocks of our prompts. The system prompt contains one of 100 personas (which consist of a few short statements about a fictional person) as well as the request to answer with a single letter or number. The system prompt in our experiments with empty personas only contained the request to answer with a single letter or number. The main prompt contains the test instruction from the BFI 2 [22] (*"Please indicate the extent to which you agree or disagree with the following statement: "I am someone who " "*.) The test instruction is directly followed by the item. After the items, we list the answer codes and the corresponding description of that code (answer) from the BFI 2. The prompt ends with the ending *Answer:*. For the "in-context" experiments, we append the previous questions with the previous answers before the new item.
For GPT-3.5 and GPT-4 we used the OpenAI chat API. We set temperature to zero, max_tokens to one and

```
messages = [{"role": "system", "content":
system_instruction},
{"role": "user","content": survey_item}]
```

where system_instruction represents the system prompt and survey_item represents the prompt. We then record the answered token if it matches one of the answer code tokens, and a non-response otherwise. In our analysis, we map non-responses to the score $s(x) = 3$.

For LLaMA2, we use the template provided in `https://huggingface.co/blog/llama2` to separate the system instructions from the prompt:

[8] `https://huggingface.co/docs/transformers/model_doc/llama2`

```
<s>[INST] <<SYS>>
{{ system_instruction }}
<</SYS>>

{{ survey_item }} [/INST]
```

where again system_instruction represents the system prompt and survey_item represents the prompt. We predict the next token based on this input and apply a softmax to the corresponding logits l to obtain probabilities p'. We then collect the subset of tokens $\{p'_t, t \in T\}$ that corresponds to the answer code tokens t and renormalize to obtain $p_t = \frac{p'_t}{\sum_{j \in T} p'_j}$. We attempted to conduct our experiments without personas and without seeding the first answer. For GPT-3.5 and GPT-4 we tried temperatures 0, 0.5 and 1. All three settings yielded no variance in most of the item responses.

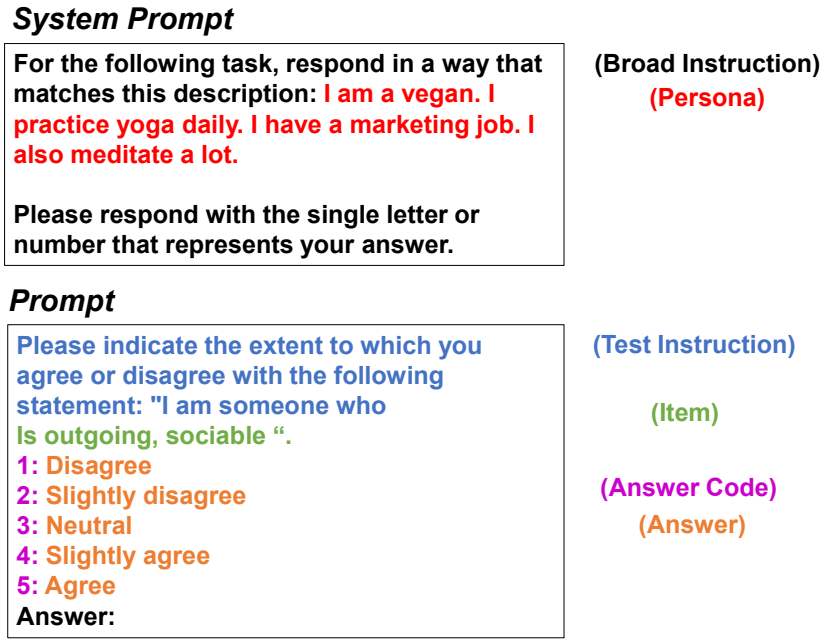

Figure 5: Prompt Template

D Personas

As [3], we use 100 personas from the PersonaChat Dataset [37]. "Empty persona prompts" simply have no persona in the system prompt and no instruction to respond according to a persona. The full list of personas can be found in our github repository **obfuscated for double blind peer review. code and data has bin submitted as .zip file**

E Details PCA

For the PCA, we used the *psych* package in R[9]. We applied varimax rotations in all PCA's. Varimax is an orthogonal rotation which was also used in the analysis of the original BFI 2 paper [22]. In summary, the PCA command in R is

[9] https://personality-project.org/r/psych/help/principal.html

```
ncomp <- 5
pca_rotated <- psych::principal(response_data,
rotate="varimax", nfactors=ncomp, scores=TRUE)
```

with ncomp defining the number of components we are looking for. In the case of personality we set this value to five in order to retrieve the five facets of personality according to the theory of the BFI 2. Some experimental setups lead to zero variance of some item responses over all $n = 100$ prompts. Therefore, the following items has been dropped for the visualization of the PCA and for the CFA.

- A37(-1) - GPT-3.5 personas in-context
- A7,E56 - GPT-3.5 in-context seeded
- A7 - GPT-4 in-context seeded

Even though this alters the structural models a little, it made it possible to compare the results.

F Details CFA

For the CFA we used the *lavaan* package in R[10]. We used the following sturctural equation models for our analysis (examples for extraversion):

Single-Component

```
extraversion <- 'E =~ E0 + E1 + E2 + E3 + E4
    + E5 + E6 + E7 + E8 + E9 + E10 + E11'
```

Defining a single extraversion component on which all 12 extraversion items load.

3 Sub-Components

```
extraversion <- 'Sociability =~ E0 + E1 + E2 + E3
                Assertiveness =~ E4 + E6 + E7 + E5
                EnergyLevel =~ E8 + E9 + E10 + E11
                Sociability ~~ 0*Assertiveness
                Sociability ~~ 0*EnergyLevel
                Assertiveness ~~0*EnergyLevel'
```

Defining three uncorrelated sub-components with 4 items each.

3 Sub-Components with 1 acquiescence factor

```
extraversion <- 'Sociability =~ E0 + E1 + E2 + E3
                Assertiveness =~ E4 + E5 + E6 + E7
                EnergyLevel =~ E8 + E9 + E10 + E11
                general_factor =~ E0 + E1 + E2 + E3
                + E4 + E5 + E6 + E7 + E8 + E9 + E10
                + E11
                Sociability ~~ 0*Assertiveness
                Sociability ~~ 0*EnergyLevel
                Assertiveness ~~0*EnergyLevel
                Sociability ~~ 0*general_factor
                Assertiveness ~~ 0*general_factor
                EnergyLevel ~~ 0*general_factor
```

Defining 3 uncorrelated sub-components with 4 items each and a general factor (acquiescence) which accounts for participants' tendency to agree/disagree. This is the model of the BFI 2 for which fit indices reach acceptable levels. However, in our experiments with LLMs, lavaan did not find a solution.

[10]https://lavaan.ugent.be/

Full Model

```
full_model <- '
              E =~ E0 + E1 + E2 + E3 + E4 + E5
              + E6 + E7 + E8 + E9 + E10 + E11
              A =~ A0 + A1 + A2 + A3 + A4 + A5
              + A6 + A7 + A8 + A9 + A10 + A11
              C =~ C0 + C1 + C2 + C3 + C4 + C5
              + C6 + C7 + C8 + C9 + C10 + C11
              N =~ N0 + N1 + N2 + N3 + N4 + N5
              + N6 + N7 + N8 + N9 + N10 + N11
              O =~ O0 + O1 + O2 + O3 + O4 + O5
              + O6 + O7 + O8 + O9 + O10 + O11'
```

This model aims at quantifying the results of our PCA. Defining 5 factor. However, in our PCA we used uncorrelated components (varimax rotated). The CFA with uncorrelated factors of the full models did not converge for the LLM data which is why we used this simpler model.
We uploaded all our R scripts and data to our github repository **obfuscated for double blind peer review. code and data has bin submitted as .zip file**

G Results CFA seeded

Table 2 shows the fit indices and reliability scores of the seeded experiment. The NA values are a result of nonconvergence of the CFA.

			Single Component			**3 Sub-Components**			**Full Model**		
Model	α	ω_h	CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA
Llama 2	**0.96**	**0.95**	0.39	0.25	0.15	NA	NA	NA	NA	NA	NA
GPT-3.5	0.67	0.69	0.74	0.68	0.10	NA	NA	NA	0.35	0.32	0.10
GPT-4	0.65	0.64	0.63	0.55	0.08	NA	NA	NA	0.33	0.30	0.10

Table 2: Reliability scores (α, ω_h) and fit indices (CFI, TLI, RMSEA) per LLM (empty personas and seeded first answer). Values on the left side of the forward slash are in-context results, on the right side are no-context results. Values are averaged over all personality traits for the Single Component and 3 Component model. Acceptable scores are bolded. NA values could not be calculated due to non-convergence or invalid solutions. Note that all reliability scores are not meaningful due to poor model fit, even though the scores are in an acceptable range.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract are a short description of our experimental results, which we discuss in detail in the paper. We motivate and theoretically ground the concerns expressed in the abstract based on established literature of psychology, measurement theory, and computer science.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a section dedicated to limitations and broader impacts.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not include theoretical results. All equations are numbered.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose all information required to reproduce our results. We release the code for our experiments, data cleaning and result analysis on github. We also release all LLM responses and other data of our experiments. We additionally explain the details of our data analysis and hardware that we used for our experiments in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As described in point 4. we release all our code and data required to reproduce our experiments. We also release the code we used for our data analysis. We provide additional instructions for reproducibility of experiments and analysis in the appendix.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all required LLM model versions and hyperparameters in the appendix. We specify the exact structural equation models of our analysis in appendix. All necessary information to understand our experiments are included in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report p-values for our first experiment and established and common fit indices for the exploratory and confirmatory factor analysis. We report thresholds for "good" fits.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the hardware we used and runtime estimates with that hardware in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We did not conduct a user study in this paper, nor do we release personal data.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in a dedicated section in the main paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Does not apply to this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite creators and owners of everything that is not original work of ours.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets introduced in this work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.