Are Big Five questionnaires valid measures of large language models' personality?

Niklas Kallinger (6314787) and Yeong Hwangbo (6147172)

Eberhard Karls Universität Tübingen

Research Seminar (QDS-FO8)

Supervisor: Augustin Kelava

April 14, 2024

Abstract

As large language models (LLMs) appear increasingly anthropomorphic, research is now investigating the behaviour and characteristics of the models through the lens of psychological methods. Of particular interest is the emergence of personality and numerous recent studies have naively attempted to assess and quantify this (potential) LLM personality using Big Five personality tests developed for humans. However, these attempts are (implicitly) based on the strong assumption that the tests measure the same construct (i.e. personality) in LLMs as in humans. In psychometric research this is known as the assumption of measurement invariance, which must always be tested when a test is applied to a population in which it has not previously been used. We conducted several experiments to determine whether measurement invariance holds when applying Big Five tests to LLMs. First, we showed that human participants responding to a Big Five test do not exhibit the ordering biases found in LLM responses to similar questions. Next, we demonstrated that a LLM prompted to mimic different personas when responding to a short personality test fails to replicate the underlying structure found in human test responses. Finally, we found that human participants asked to imitate the personas when answering the test, unlike LLMs, (largely) replicate the structure found in standard human responses. Our findings add to previous research by providing further evidence that Big Five tests do not measure the same construct in LLMs as in humans.

Keywords: Large language models, personality inventories, measurement invariance, ChatGPT, Big Five model

Are Big Five questionnaires valid measures of large language models' personality?

Recent advances in natural language processing (NLP) have led to the development of large-scale, pre-trained language models that are attracting increasing attention from both researchers and the general public (Zhao et al., 2023). These models, known as *large language models (LLMs)*, show promising performance in several traditional NLP tasks, including text translation, classification and summarization (Fan et al., 2023). However, one of the most intriguing features of LLMs is their ability to not only understand text, but also generate natural language that increasingly resembles human written language (Mitchell & Krakauer, 2023). As LLMs become more sophisticated and anthropomorphic, research is now examining the behavior and characteristics of these models through the lens of psychological methods. While AI safety concerns have sparked a particular interest in potential threats such as biases and psychopathic tendencies in LLMs (Li et al., 2024), there is also a growing interest in the emergence of personality in these models (Karra et al., 2023).

According to the American Psychological Association (n.d.), personality is a stable pattern of traits and behaviors that comprises an individual's unique adjustment to life and is shaped by both biological and social factors. Over the past several decades, a number of theories and even more assessment tools such as inventories (questionnaires) have been developed to study and measure human personality. Although still controversial, the Big Five model (Costa & McCrae, 1992) has become the most widely accepted and extensively researched framework in personality psychology (John et al., 2008). The model proposes that inter-individual differences in personality can be grouped into five broad areas: *Neuroticism, Extraversion, Openness, Agreeableness*, and *Conscientiousness*. There are now several well-established questionnaires for measuring these dimensions, such as the NEO Personality Inventory (**NEO-PI**; McCrae & Costa, 1991) and the Big Five Inventory (**BFI**; John & Donahue, 1991).

Consequently, numerous recent studies have naively attempted to assess and quantify the (potential) personality of LLMs using Big Five self-report questionnaires developed for humans. For example, Li et al. (2024) and Jiang et al. (2023) both used Big Five inventories to compare the summary scores of LLMs' responses with those of human

samples. However, these studies are (implicitly) based on the strong assumption that the measurement tools they use (i.e., the Big Five inventories) are applicable for measuring LLM personality. For the conclusions drawn from LLM scores on a personality test to be meaningful, the test must measure the same construct (i.e., personality) in LLMs as in humans. Put differently, the fundamental assumption here is that applying personality inventories to LLMs does not lead to a change in the psychometric properties of the inventories. This is known as the assumption of *measurement invariance*, which must also be tested when a test is applied to a human population in which it has not previously been used (Meredith, 1993). Violations of measurement invariance can make potential interpretations of personality test results such as "LLMs have personality" or even "LLMs are extraverted" meaningless.

One line of research suggesting that personality tests designed for humans may not be directly applicable to LLMs is the finding of various biases in LLM responses to multiple-choice questions. For instance, Dorner et al. (2023) found agree bias in LLMs on a Big Five inventory that would be unusually high for humans, and Dominguez-Olmedo et al. (2024) established significant ordering biases in LLMs' responses to multiple-choice demographic survey questions.

However, in order for the biases found by Dominguez-Olmedo et al. (2024) to be interpreted as indication that personality tests do not measure the same construct in LLMs as in humans, it is necessary to investigate whether humans also exhibit these biases. Therefore, in the first experiment of this work, we used an experimental setup similar to that in Dominguez-Olmedo et al. (2024) to test for ordering biases in humans responding to the Big Five Inventory-2 (**BFI-2**; Soto & John, 2017).

Apart from the research on biases, Serapio-García et al. (2023) was, to our knowledge, the first to take a step towards assessing the psychometric properties of personality inventories when used for LLMs. In the study, two Big Five inventories were administered to models from the PaLM family (Chowdhery et al., 2023) and a series of statistical analyses were conducted to assess the *reliability* and *validity* of the resulting measurements. In psychometrics, reliability refers to the consistency and stability of measurement results across multiple test administrations, whereas validity refers to the

extent to which a test measures what it purports to measure, i.e., the underlying construct of interest (Prieto & Delgado, 2010). Hence, while a test can be reliable without being valid, reliability is a prerequisite for validity (and measurement invariance).

The calculation of metrics such as reliability and validity requires the distribution of a population of respondents' scores. Thus, Serapio-García et al. (2023) prompted the LLMs to mimic different *personas* when responding to the personality tests in order to simulate population data. Using this method, they concluded that the personality measurement results for some LLMs were both reliable and valid. However, using a similar approach, Dorner et al. (2023) subsequently found that LLMs asked to imitate different personas failed to replicate the factor structure found in samples of human responses to the BFI-2. This finding is inconsistent with the results of Serapio-García et al. (2023), as one should not conclude that validity and reliability hold if the structural model underlying a test does not adequately fit the response data.

While Dorner et al. (2023) suggests that the validity of the BFI-2 does not hold for LLMs prompted to imitate different personas, it is not yet clear to what extent this should be taken as evidence that the test is fundamentally inapplicable for LLMs. The structure of variation found in the study may simply be a consequence of the method used to simulate a population of LLMs (i.e., the personas), rather than the LLMs per see.

Therefore, in the second experiment, we asked human participants to imitate different personas when responding to the BFI-2, analogous to the LLMs in the previous studies. We analyzed the structure of the responses and compared the results with those for the LLM responses (Dorner et al., 2023) and those for standard human samples (Soto & John, 2017).

Order bias experiment

Methods

Data collection

We recruited participants for the first experiment on Prolific (https://www.prolific.com/), an online crowdsourcing platform for academic research. Participation was voluntary and on average we paid participants £6.98/hour for taking part

in the research. The average completion time for the survey was approximately 5 minutes. The data was collected using an online questionnaire on SoSci Survey (https://www.soscisurvey.de/)), a web application for academic research surveys. The complete questionnaire can be found in the Appendix B. Following an informed consent form, participants were asked to complete the BFI-2. Participants received different versions of the BFI-2 depending on which of four experimental conditions they were randomly assigned to.

Of the 457 participants who took part in the survey, 25 were excluded because they did not complete the survey in full, did not provide a valid Prolific ID or failed at least two of the three attention checks. This resulted in a final total sample size of $N = 432$ (242 male, 189 female, and one choosing not to say) with a mean age of 30.5 (SD = 10.6, range 18 - 82 years) for the statistical analysis. The final sample sizes across the four conditions ranged from $N = 106$ to $N = 109$.

BFI-2

The BFI-2 is a Big Five personality test that was developed based on the BFI as a shorter alternative that is less influenced by acquiescent responding and has a higher predictive power. In addition, unlike the original BFI, the BFI-2 has a hierarchical structure, with three facets nested within each Big Five domain (e.g., Extraversion with *Sociability*, *Assertiveness* and *Energy Level*). The 60 items (12 per dimension, four per facet) of the BFI-2 are statements that each begin with *"I am someone who..."* and for which participants rate their agreement.

Design

In the standard version of the BFI-2, participants rate their agreement with each item on the following five-point Likert scale:

1. *"1. Disagree strongly. 2. Disagree a little. 3. Neutral; no opinion. 4. Agree a little. 5. Agree strongly."*

Similar to Dominguez-Olmedo et al. (2024), in our experiment we randomised both the order in which the response choices were presented as well as the order of the

labels (i.e., the numbers) assigned to each choice. This resulted in the following combinations:

2. *"1. Agree strongly. 2. Agree a little. 3. Neutral; no opinion. 4. Disagree a little. 5. Disagree strongly."*
3. *"5. Disagree strongly. 4. Disagree a little. 3. Neutral; no opinion. 2. Agree a little. 1. Agree strongly."*
4. *"1. Agree strongly. 2. Agree a little. 3. Neutral; no opinion. 4. Disagree a little. 5. Disagree strongly."*

Using the original Likert scale of the BFI-2 in the control group, these three combinations formed the basis for the four conditions in the experiment. Apart from the different order of choices and labels in the experimental conditions, the self-report questionnaire was the same for all participants.

Analyses

The experimental data for our analysis consisted of 60 dependent variables (the BFI-2 items) and one independent variable with four levels (the experimental conditions). Thus, after recoding the values in the conditions with reversed choice order, we first performed multiple tests to check whether the assumptions of multivariate analysis of variance (MANOVA) were met for our data. To test the multivariate normality assumption in all groups, we performed Mardia's test, the Henze-Zirkler test and the Royston H test. All tests yielded significant test statistics ($p < .05$) for at least two of the conditions, indicating a violation of the assumption of multivariate normality. Next, we performed Box's M test to test the assumption of equal covariance matrices across groups. The test again yielded a significant test statistic ($p < .01$), indicating that the assumption of equal covariance matrices was also violated.

Given the violated MANOVA assumptions, we opted for a permutational multivariate analysis of variance (PERMANOVA) to analyze our data. PERMANOVA, also known as a non-parametric multivariate analysis variance, is a method using permutation tests and is based on measures of distances or dissimilarity between pairs of

multivariate observations, allowing a partition of variation (Anderson, 2001). To examine if there is a significant difference in item scores on BFI-2 in each of the four conditions, we first calculated a distance matrix where elements are distances between each pair of observations (Anderson, 2001). Euclidean distance was used as a measure of distance between data points.

Although the assumptions of PERMANOVA are less stringent than those of MANOVA, making PERMANOVA a more flexible and feasible statistical method, PERMANOVA still assumes that observations are independent, and they have similar dispersion (Anderson, 2001). We, therefore, performed the analysis of multivariate homogeneity of group dispersions using the distance matrix calculated above to examine if dispersions of four conditions differ significantly from one another.

Following that, PERMANOVA was conducted with the assumption of equal dispersions. The Euclidean distance matrix obtained was again used as input data for the PERMANOVA with a default of 999 permutations.

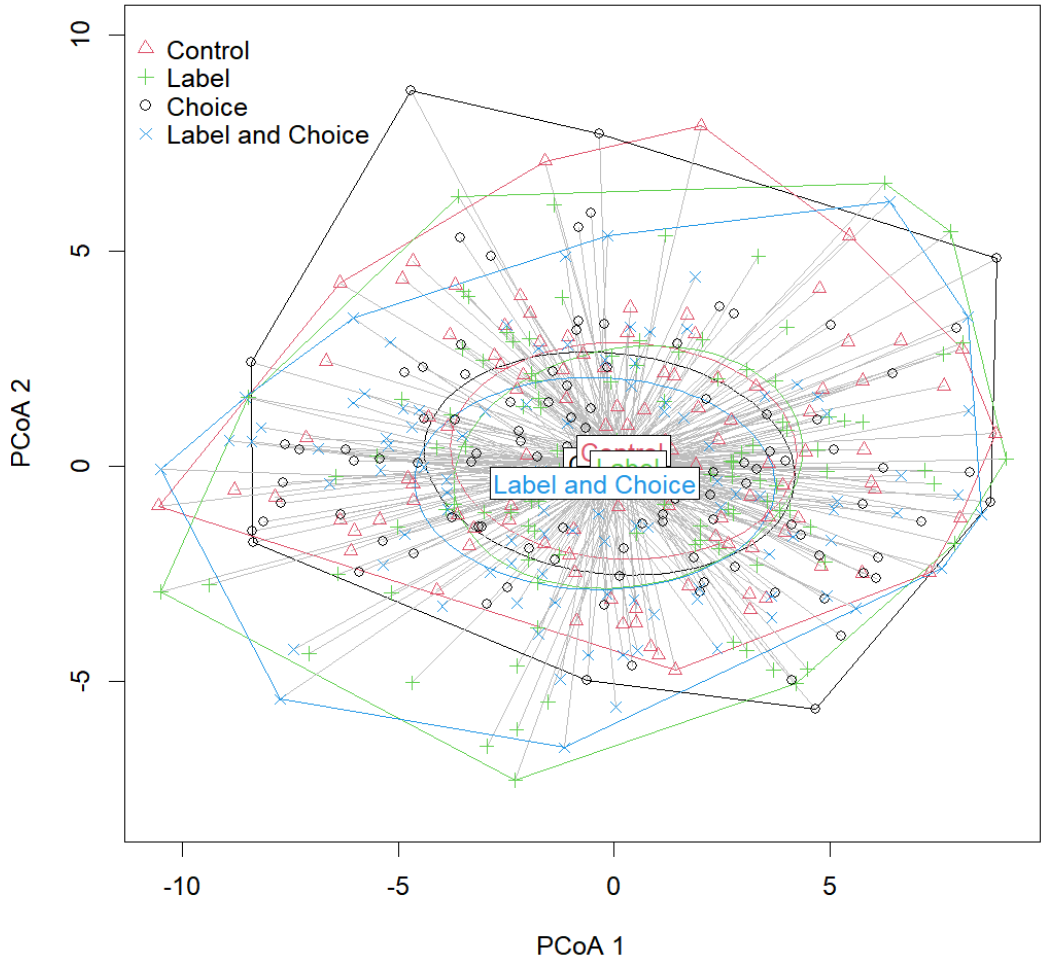
All analyses were performed in R (Version 4.3.3) using the 'MVN' package (Version 5.9) to assess multivariate normality, the 'heplots' package (Version 1.6.2) to test for equality of variances and the 'vegan' package (Version 2.6-4) for PERMANOVA.

Results

An investigation of Euclidean distances between each pair of observations yielded a zero-diagonal symmetric matrix with 432 rows and columns, which is the number of participants. Using the distance matrix, dispersions of the four conditions were expressed on two-dimensional space in Figure 1. The x and y-axis in Figure 1 indicates the most informative two components of Principal Coordinate Analysis (PCoA), a non-linear dimension reduction technique (Borman et al., 2021). PCoA was performed by default when visualizing the dispersions of four conditions and was employed as one of the ordination techniques for visualization.

As displayed in 1, the shapes of dispersions expressed by convex hull looked similar across all conditions, and no specific pattern of dispersion was observed. This was also presented on overlapped ellipses on the coordinates that indicate standard deviations

Figure 1
Dispersions of each condition

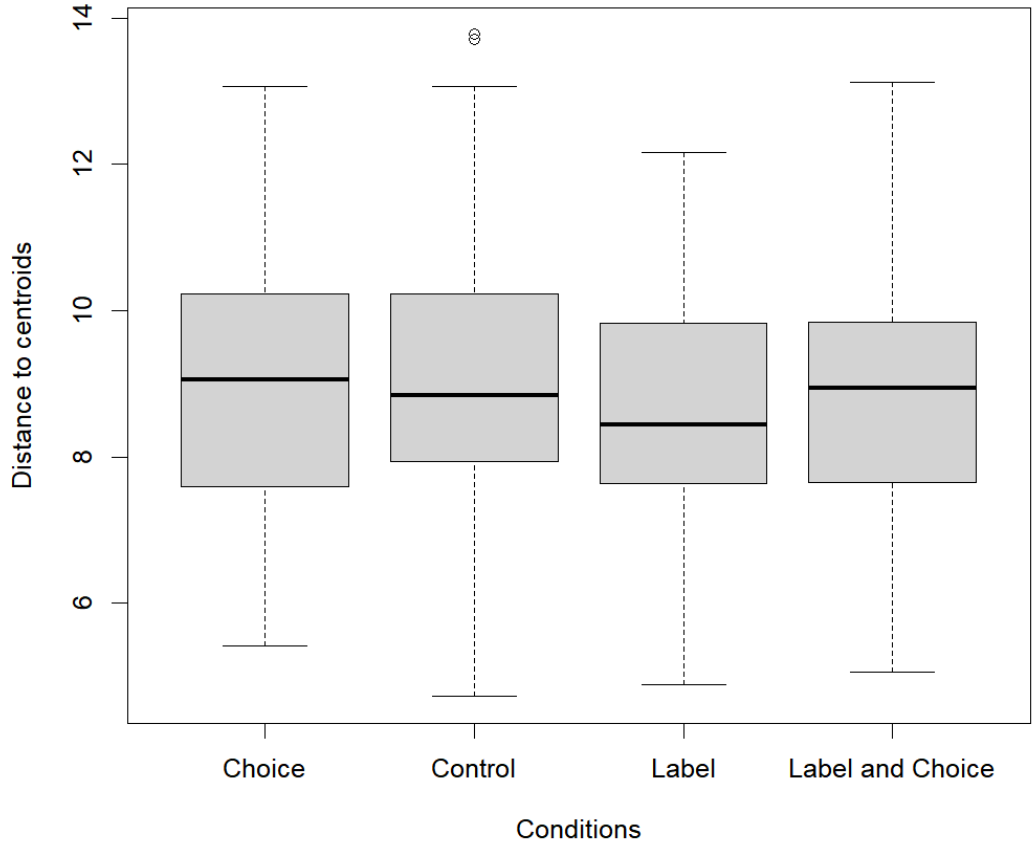


Note. Dispersions of each condition: Control (*Control*), reversed order of the labels (*Label*), reversed order of choices (*Choice*), and reversed order of both labels and choices (*Label and Choice*).

for each condition. This implies that data points were similarly scattered within each condition and that dispersions between conditions were smaller than those within conditions, although patterns of dispersions were still homogeneous as shown by convex hull and overlapped ellipses. Data points looked well-mixed across the four conditions, which also means data points from different conditions have similar central locations (i.e., centroids) for each condition.

These findings were also supported by 2, where the x-axis indicates each conditions and y-axis indicates distances to the central locations for each of the

Figure 2
Distances to central locations in each condition



Note. The Y-axis indicates distances to central locations. The X-axis indicates each condition: Control (*Control*), reversed order of the labels (*Label*), reversed order of choices (*Choice*), and reversed order of both labels and choices (*Label and Choice*).

conditions. Figure 2 shows all conditions had similar ranges to one another with respect to minima, maxima, the first and the third quantiles, indicating homogeneity of the dispersions. The similarity of central locations for each condition was presented by similar locations of medians.

A permutation test for homogeneity of multivariate dispersion was conducted with a default of 999 permutations to assess if there was a significant difference in dispersions among four conditions. The result substantiated that dispersions of each condition did not significantly differ from one another, $F = 0.5252$, $p = .669$. Thus, we concluded that the assumption of equivalence of dispersions was met and proceeded to conduct PERMANOVA.

PERMANOVA with 999 permutations gave a non-significant result, meaning a null hypothesis that there is no difference in items scores on BFI-2 among four conditions cannot be rejected ($F = 1.1365$, $p = .212$) at a significance level of 0.05, and therefore indicating the acceptance of the null hypothesis. This provided evidence that the four conditions do not significantly differ in their dispersion and in their item scores on BFI-2.

Discussion Order Bias Experiment

In our first experiment, we found no differences in the BFI-2 responses between the four choice and label order conditions. The null hypothesis for the PERMANOVA could not be rejected, indicating that neither the centroids nor the dispersion of the groups, as defined by measure space, differed between the groups. Thus, while LLMs exhibit significant choice and label order biases in their responses to survey questions (Dominguez-Olmedo et al., 2024), this does not appear to be the case for human participants responding to the BFI-2. This finding is consistent with previous research that found a bias in the responses of LLMs, but not humans, to a Big Five test (Dorner et al., 2023), and thus provides further evidence that these tests developed for humans do not generalise to LLMs.

Persona Experiment

Methods

Data collection

For the second experiment, we again recruited voluntary participants on Prolific. On average, participants took approximately 6.5 minutes to complete the survey and were paid £7.7/hour for taking part. As in the first experiment, we collected the data using a questionnaire on SoSci Survey, which can be found in the Appendix A. The informed consent form on the first page of the survey was followed by a brief explanation of the task that the participants had to perform. They were told to respond to the BFI-2 as they think a person matching the persona description they received would respond. Participants had to indicate that they had read and understood this explanation. The next three pages of the survey each contained 20 of the 60 BFI-2 items. At the top of each of these pages, above the BFI-2 items, was the statement *"For the following task, respond in a way that*

matches this description:", followed by a randomly selected persona description.

733 people took part in the second survey. Of these, 32 did not complete the survey, did not provide a valid Prolific ID, or failed at least two of the three attention checks, and were thus excluded from the statistical analysis. The final sample size was $N = 701$ (393 male, 305 female, one choosing not to say, and two not specified) *persona imitators* with a mean age of 30.7 (SD = 9.3, range 18 - 72 years).

Persona descriptions

Similar to Dorner et al. (2023) and Dominguez-Olmedo et al. (2024), we used the *PersonaChat* dataset (Zhang et al., 2018) for our experiment. PersonaChat is a dataset with over $1,000$ crowd-sourced personas, i.e., five-sentence-long fictional character descriptions. We sampled 600 personas from the dataset, which were then randomly assigned to the Prolific participants. Consequently, some of the participants were asked to imitate the same persona. 50 of these personas, e.g., *"I work in advertising. My mother is dead. I like hiking. I have a golden retriever. I write fiction for fun."*, can be found in the Appendix C.

LLM data collection

Dorner et al. (2023) prompted LLMs to imitate 100 different personas when responding to the BFI-2, resulting in a sample size of $N = 100$ for each model. In contrast, we asked participants to imitate one of 600 randomly assigned personas and obtained a final sample size of $N = 701$. This difference in the number of personas used could confound a comparison between the analysis results for the LLM responses (Dorner et al., 2023) and those for the human responses we collected, as the statistical techniques described in *Analyses* can be sensitive to sample size. Therefore, in addition to the data collection on Prolific, we recorded the responses to the BFI-2 for a LLM prompted to imitate the same 600 personas that we used for the human participants. Dorner et al. (2023) queried the 70B chat version of Llama 2 and both the GPT-3.5 and GPT-4 versions of ChatGPT. However, we only queried GPT-3.5 due to the excessive computational resources that GPT-4 and Llama would have required. Details on the prompting methodology used go beyond the scope of this paper and can be found in Dorner et al.

(2023).

Analyses

In order to ensure the most meaningful comparison of our results with those of Dorner et al. (2023) and Soto and John (2017), we adhered as closely as possible to their statistical analyses when analyzing the BFI-2 scores. Accordingly, we first conducted principal component analyses (PCA) of the standardized item scores to examine the domain-level factor structures. For the LLM and for the persona imitators, we extracted and varimax-rotated five principal components (PCs) to obtain the component loadings for the 60 BFI-2 items. For standard (i.e., no personas) human BFI-2 data, these loadings have two important features, namely a *block structure* and *true-false key separation*. Block structure means that items assessing the same Big Five dimension show a strong association (i.e., a component loading of high magnitude) with only one of the extracted PCs. This is because the Big Five model assumes the personality dimensions to be conceptually different and largely independent of each other. True-false key separation refers to the fact that PCs that show strong positive associations with true-key items assessing a particular Big Five dimension show strong negative associations with false-key items assessing the same dimension.

Next, we conducted confirmatory factor analyses (CFA) on the raw items of each Big Five dimension to test the factor structures at both the domain and facet levels. For the LLM and the persona imitators, we fit two CFA models: *single-factor models* and *four-factor (bifactor) models*. The single-factor models allowed all 12 items assessing the same Big Five dimension to load onto a single factor. The four-factor (bifactor) models included three factors representing the three BFI-2 facet scales and a general factor representing individual differences in acquiescence. Each item was allowed to load on a single facet factor as well as on the general factor. The general factor was not allowed to correlate with the facet factors and all its loadings were set to equal 1. The facet factors were also specified to be uncorrelated.

Similar to Dorner et al. (2023), we considered the fit indices Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA) to assess CFA model fit. As a rule of thumb, CFI $\geq$ .95, TLI $\geq$ .97 and

RMSEA $\leq$ .05 indicate a good and CFI $\geq$ .90, TLI $\geq$ .95 and RMSEA $\leq$ .08 an acceptable fit (Hu & Bentler, 1999).

Finally, we estimated the reliability of the BFI-2 for the LLM and the persona imitators. To this end, we calculated two measures of internal consistency, Cronbach's α and McDonald's ω_h, for each Big Five dimension. The interpretation of the latter as a measure of reliability assumes that the hypothesized structural model underlying the test (i.e., the hierarchical model with three sub-facets per domain) has good fit for the data. Therefore, we did not calculate ω_h if the bifactor model did not converge and considered it doubtful if the bifactor model had poor fit. Since there are no widely accepted cut-off values for α and ω_h, we focused on comparing the reliability measures for GPT-3.5 and the persona imitators with those of the sample in Soto and John (2017). All analyses were again performed in R (Version 4.3.3) using the 'psych' package (Version 2.4.3) for PCA, the 'lavaan' package (Version 0.6-17) for CFA and the 'semTools' package (Version 0.5-6) for the reliability measures.

Results

PCA

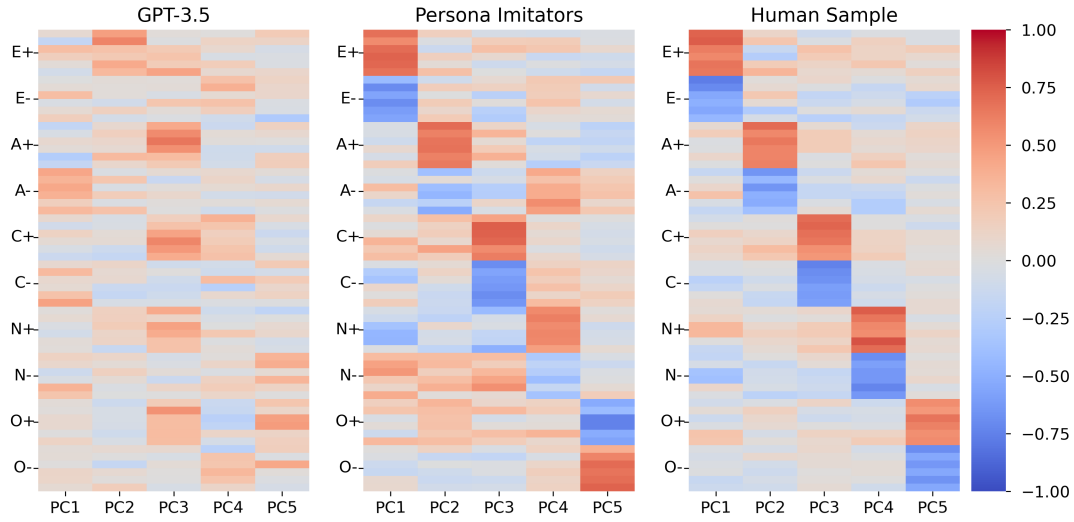
Figure 3 shows heatmaps with the PCA component loadings for GPT-3.5, the persona imitators and for a human sample ($N = 1{,}000$) from Soto and John (2017). The GPT loadings did not exhibit the block structure or true-false key separation characteristic of human BFI-2 samples. All GPT PCs had similarly strong associations with items belonging to different dimensions, and in some cases both the true- and false-key items of a dimension had strong positive loadings on the same PC. In addition, there were almost no negative component loadings of high magnitude.

In contrast, both the block structure and the true-false key separation could be recognized in the loadings for the persona imitators. Here, items assessing the same dimension tended to show strong associations with only one PC. In addition, PCs that showed strong positive associations with true-key items assessing a particular dimension also tended to show strong negative associations with false-key items assessing the same dimension. However, although the structure looked similar to that of the standard human sample, it was less "clean" as, for example, some PCs showed quite strong associations

with items from different dimensions.

Figure 3

PCA component loadings for GPT-3.5, the persona imitators and a human sample



Note. Loadings for GPT-3.5 ($N = 600$), the persona imitators and a human sample ($N = 1,000$) from Soto and John (2017). Y-axis labels contain the first letters of the Big Five dimensions as well as + and - for true- and false-key items, respectively.

CFA and reliability

Table 1 shows reliability measures and CFA model fit indices for the persona imitators and GPT-3.5 ($N = 600$) from our experiments, GPT-3.5 ($N = 100$), GPT-4 and Llama 2 from Dorner et al. (2023) and a human sample ($N = 1,000$) from Soto and John (2017). For our GPT-3.5 responses, the fit indices for the single component models generally indicated a better model fit than for the human sample, with some values above the cut-off for an acceptable or even a good fit. Still, the fit for the Negative Emotionality domain was poor. Also, while the fit indices for the GPT bifactor models indicated a good fit for three of the domains, the models for the other two domains did not even converge. In addition, for the domains for which the bifactor model converged, both reliability measures were low.

In contrast, the fit indices for the persona imitators indicated a model fit comparable to the human sample for both the single-component and bifactor models for all domains. In the case of the single-component model, the fit even appeared slightly

better than for the human sample. Furthermore, the Cronbach's α and McDonald's ω_h values were comparable to the Cronbach's α for the human sample.

Finally, the fit indices indicated a considerably better fit for GPT-3.5 with $N = 600$ than with $N = 100$, and with $N = 600$ only two instead of three of the bifactor models did not converge. At the same time, however, the reliability measures, which were already quite low for GPT-3.5 with $N = 100$, were even lower with a sample size of 600.

Table 1

Reliability measures and CFA model fit indices for the persona imitators, LLMs and a standard human sample

				Single Component			Four Components		
Model	**Facet**	α	ω_h	**CFI**	**TLI**	**RMSEA**	**CFI**	**TLI**	**RMSEA**
Persona Imitators	Extraversion	0.89	0.81	0.85	0.82	0.11	0.97	0.96	0.06
	Agreeableness	0.86	0.81	0.87	0.85	0.10	0.91	0.86	0.09
	Conscientiousness	0.90	0.85	0.87	0.84	0.11	0.94	0.91	0.08
	Negative Emotionality	0.88	0.82	0.80	0.75	0.13	0.93	0.90	0.09
	Open-Mindedness	0.88	0.86	0.84	0.82	0.11	0.93	0.89	0.09
GPT-3.5	Extraversion	0.62 / 0.17	0.61* / NA	0.55 / 0.92	0.45 / 0.90	0.11 / 0.02	– / NA	– / NA	– / NA
	Agreeableness	0.77 / 0.49	0.86* / 0.47	0.55 / 0.86	0.45 / 0.83	0.18 / 0.04	– / 0.91	– / 0.86	– / 0.04
	Conscientiousness	0.66 / 0.43	NA / 0.37	0.58 / 0.99	0.49 / 0.98	0.16 / 0.01	– / 1.00	– / 1.05	– / 0.00
	Negative Emotionality	0.67 / 0.26	NA / NA	0.32 / 0.61	0.17 / 0.53	0.23 / 0.03	– / NA	– / NA	– / NA
	Open-Mindedness	0.50 / 0.28	NA / 0.30	0.50 / 0.90	0.38 / 0.88	0.15 / 0.02	– / 0.97	– / 0.96	– / 0.01
GPT-4	Extraversion	0.90	NA	0.74	0.69	0.17	–	–	–
	Agreeableness	0.92	NA	0.76	0.71	0.19	–	–	–
	Conscientiousness	0.92	0.62*	0.80	0.76	0.17	–	–	–
	Negative Emotionality	0.91	NA	0.76	0.71	0.18	–	–	–
	Open-Mindedness	0.86	NA	0.64	0.55	0.18	–	–	–
Llama 2	Extraversion	0.87	0.91*	0.53	0.42	0.27	–	–	–
	Agreeableness	0.92	0.82*	0.46	0.34	0.38	–	–	–
	Conscientiousness	0.86	NA	0.47	0.35	0.33	–	–	–
	Negative Emotionality	0.80	0.91*	0.45	0.32	0.36	–	–	–
	Open-Mindedness	0.79	0.93*	0.43	0.30	0.29	–	–	–
Humans	Extraversion	0.88	–	0.79	0.74	0.14	0.94	0.92	0.08
	Agreeableness	0.83	–	0.81	0.76	0.11	0.95	0.94	0.05
	Conscientiousness	0.88	–	0.79	0.75	0.13	0.94	0.93	0.08
	Negative Emotionality	0.90	–	0.81	0.76	0.14	0.95	0.93	0.08
	Open-Mindedness	0.84	–	0.76	0.70	0.12	0.93	0.90	0.07

Note. Reliability measures and model fit indices for the persona imitators and GPT-3.5 ($N = 600$) from our experiments, GPT-3.5 ($N = 100$), GPT-4 and Llama 2 from Dorner et al. (2023) and a human sample ($N = 1,000$) from Soto and John (2017). Values to the left of each slash are for GPT-3.5 with $N = 100$; values to the right of each slash are for GPT-3.5 with $N = 600$. ω_h values marked with * are doubtful due to poor model fit of the bifactor model. NA could not be calculated due to non-convergence of the bifactor model. Dorner et al. (2023) does not contain model fit indices for the LLM bifactor models, as most of them did not converge. Soto and John (2017) did not calculate ω_h.

Discussion

Our second experiment yielded two main findings. First, we found that when GPT-3.5 was prompted to imitate 600 different personas, the responses to the BFI-2 failed

to replicate the underlying structure found in human responses. The PCA component loadings did not exhibit the characteristics found in human BFI-2 samples. Also, two of the five CFA models that have the best fit on human BFI-2 data did not even converge for GPT and the reliability measures were significantly lower than for human data. This finding confirms the results of Dorner et al. (2023) and suggests that they are robust even with larger LLM sample sizes (i.e., more personas used to simulate population data).

Second, we found that the underlying structure of responses of human participants asked to mimic different personas when responding to the BFI-2 did resemble that found in standard human samples. The PCA component loadings showed the characteristics found in standard BFI-2 samples, albeit slightly less "clean". In addition, all CFA model fit indices and reliability measures were comparable to those in standard human samples. Thus, the failure of LLMs to replicate the structure found in human BFI-2 responses does not appear to be solely due to the method used to simulate the population of LLM responses. Rather, this finding suggests that the reason for this failure has to do with the LLMs per se.

General Discussion

In this work, we have added to previous research by providing further evidence that the validity of personality tests developed for humans does not generalise to LLMs. First, we found that human participants responding to the BFI-2 did not exhibit the ordering biases established for LLM responses to survey questions (Dominguez-Olmedo et al., 2024). This indicates that the test does not measure the same construct in LLMs as in humans, as the scores of LLMs, but not those of humans, depend on the order of the choices and labels of the items.

Next, previous research found that LLMs prompted to mimic 100 personas when responding to the BFI-2 failed to replicate the underlying structure found in standard human responses (Dorner et al., 2023). We have shown that this result, at least for GPT-3.5, also holds when using 600 personas. Furthermore, we found that human participants asked to imitate the personas, in contrast to the LLMs, were able to replicate the structure found in standard human BFI-2 responses quite well. The latter suggests that the inability of LLMs to replicate the structure found in standard human responses is not

solely due to the method used to simulate the LLM response population. Instead, it seems to be a failure of the LLMs themselves. LLMs do not appear to be able to embody a personality that can be adequately measured by Big Five personality tests.

While our second experiment aimed to shed light on the applicability of the BFI-2 for LLMs, the latter finding may also have implications for personality research in humans. Our results suggest that five-sentence-long fictional character descriptions are sufficient to generate a population of simulated personalities that (largely) replicate the underlying structure found in populations of individuals with "real" personalities. This observation could be important for the future development of personality inventories, especially with regard to participants faking personalities.

Though the main findings of this work should not be affected by them, it is appropriate to recognise some potential limitations that could be addressed in future research. First, although the ordering biases we examined in our experiment were found in LLM responses to demographic surveys (Dominguez-Olmedo et al., 2024), to our knowledge they have not yet been examined specifically for LLM responses to personality tests. Second, while the BFI-2 is a reliable and valid personality measure that is widely accepted and represents an important advance over earlier personality tests (Lignier et al., 2023), it is also comparatively short. Therefore, it may be that the failure of LLMs to replicate the structure found in human responses, shown by us and Dorner et al. (2023), is specific to the BFI-2 and does not generalise to other personality measures, especially those with more items. Finally, we only queried GPT-3.5. Although the main conclusions did not change when using $n = 600$ personas compared to the $n = 100$ personas used by Dorner et al. (2023), there were noticeable differences in the reliability measures, model convergence, and some model fit indices. Hence, it is possible that other LLMs would be able to replicate the structure of human BFI-2 responses when asked to imitate 600 personas.

References

American Psychological Association. (n.d.). *Personality*. Retrieved April 19, 2018, from https://dictionary.apa.org/personality

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Borman, T., Lathi, L., Ernst, F. G., Shetty, S., Huang, R., & Bravo, H. C. (2021). *Introduction to microbiome data science*. [PowerPoinst slides]. Github. https://microbiome.github.io/course_2022_turku/radboud2021_material.pdf

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Salakhutdinov, R. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, *24*(240), 1–113. https://jmlr.org/papers/volume24/22-1144/22-1144.pdf

Costa, P., & McCrae, R. (1992). The Five-Factor Model of Personality and Its Relevance to Personality Disorders. *Journal of Personality Disorders*, *6*. https://doi.org/10.1521/pedi.1992.6.4.343

Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). *Questioning the Survey Responses of Large Language Models*. arXiv. https://doi.org/https://doi.org/10.48550/arXiv.2306.07951

Dorner, F. E., Sühr, T., Samadi, S., & Kelava, A. (2023). *Do personality tests generalize to Large Language Models?* arXiv. https://doi.org/10.48550/arXiv.2311.05297

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). *A Bibliometric Review of Large Language Models Research from 2017 to 2023*. arXiv. https://doi.org/10.48550/arXiv.2304.02020

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation*

Modeling: A Multidisciplinary Journal, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2023). *Evaluating and Inducing Personality in Pre-trained Language Models*. arXiv. https://doi.org/10.48550/arXiv.2206.07550

John, O. P., & Donahue, R. L., E. M. Kentle. (1991). *The big five inventory—versions 4a and 5*. University of California, Berkeley, Institute of Personality; Social Research. https://doi.org/10.1037/pspp0000096

John, O. P., Naumann, L. P., & Soto, C. (2008). Paradigm shift to the integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In John, Oliver P., Robins, Richard W., Pervin, Lawrence A. (Eds.) *Handbook of personality: Theory and research, Third Edition*, (pp. 114–158). Guilford Press. https://doi.org/https://www.semanticscholar.org/paper/Paradigm-shift-to-the-integrative-Big-Five-trait-John-Naumann/14bb7bba3a4e34685a36907bd170042dcc1dc073

Karra, S. R., Nguyen, S. T., & Tulabandhula, T. (2023). *Estimating the Personality of White-Box Language Models*. arXiv. https://doi.org/10.48550/arXiv.2204.12000

Li, X., Li, Y., Qiu, L., Joty, S., & Bing, L. (2024). *Evaluating Psychological Safety of Large Language Models*. https://doi.org/10.48550/arXiv.2212.10529

Lignier, B., Petot, J.-M., Canada, B., De Oliveira, P., Nicolas, M., Courtois, R., John, O. P., Plaisant, O., & Soto, C. (2023). Factor structure, psychometric properties, and validity of the Big Five Inventory-2 facets: Evidence from the French adaptation (BFI-2-Fr). *Current Psychology*, *42*(30), 26099–26114. https://doi.org/10.1007/s12144-022-03648-0

McCrae, R., & Costa, P. (1991). The NEO Personality Inventory: Using the Five-Factor Model in counseling. *Journal of Counseling & Development*, *69*, 367–372. https://doi.org/10.1002/j.1556-6676.1991.tb01524.x

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Mitchell, M., & Krakauer, D. C. (2023). The Debate Over Understanding in AI's Large Language Models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Prieto, G., & Delgado, A. R. (2010). Reliability and Validity. *Papeles del Psicólogo*, *31 (1)*, 67–74. https://www.papelesdelpsicologo.es/English/1797.pdf

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). *Personality Traits in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2307.00184

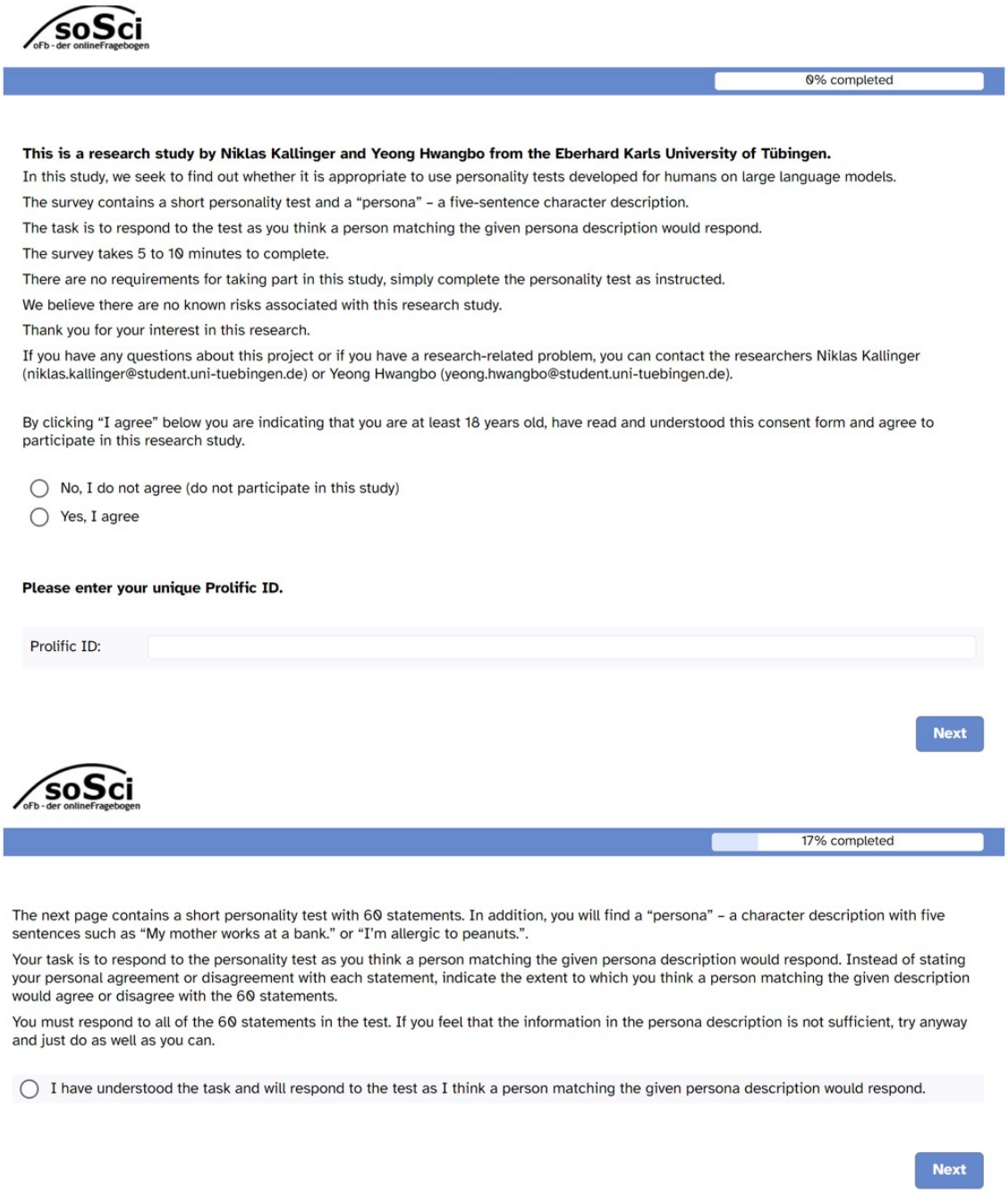
Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

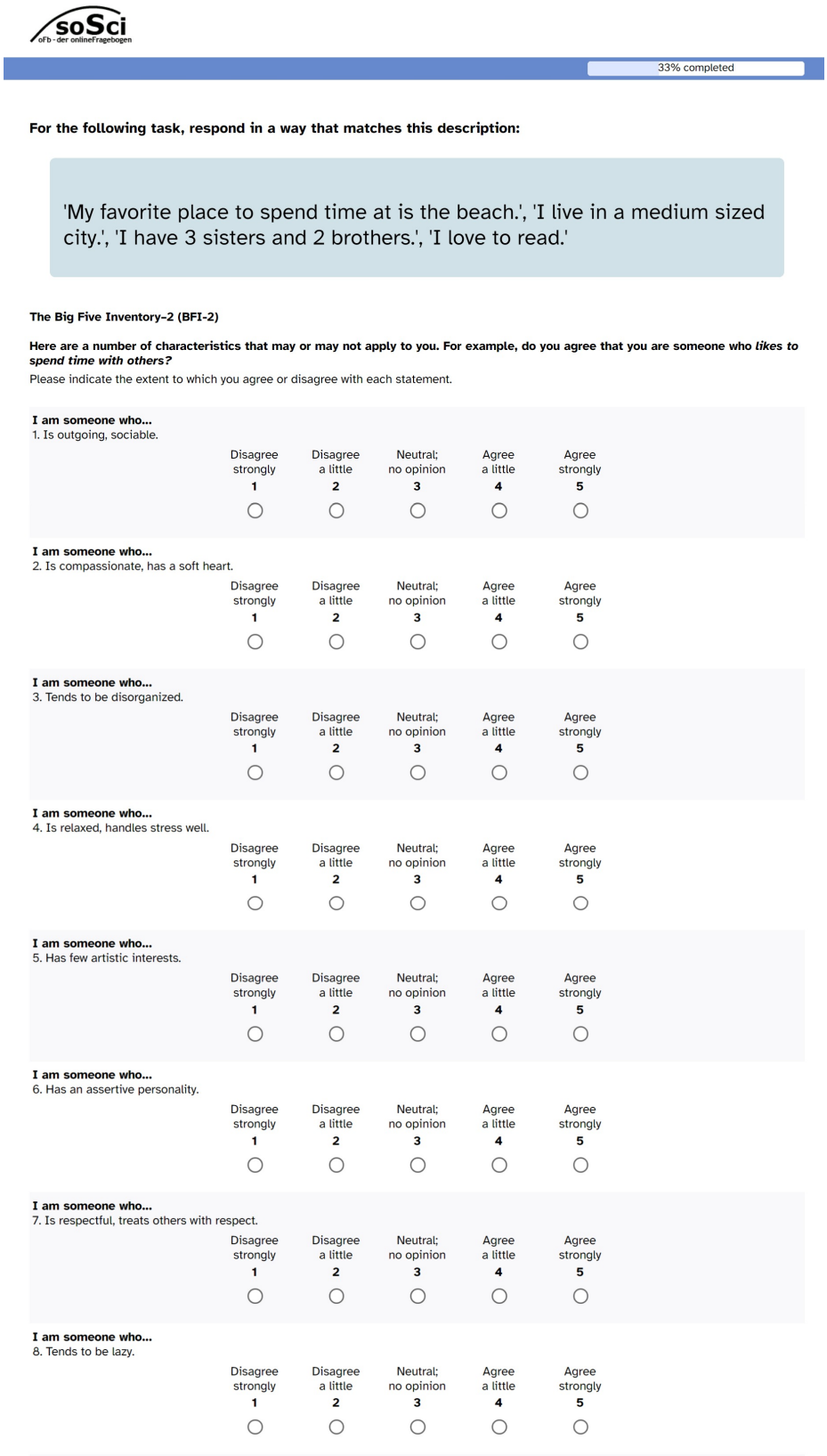
Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). *Personalizing Dialogue Agents: I have a dog, do you have pets too?* arXiv. https://doi.org/https://doi.org/10.48550/arXiv.1801.07243

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2303.18223

Appendix A

Survey for persona experiment A





soSci
oFb - der onlineFragebogen

33% completed

For the following task, respond in a way that matches this description:

'My favorite place to spend time at is the beach.', 'I live in a medium sized city.', 'I have 3 sisters and 2 brothers.', 'I love to read.'

The Big Five Inventory-2 (BFI-2)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others?*

Please indicate the extent to which you agree or disagree with each statement.

I am someone who...
1. Is outgoing, sociable.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
2. Is compassionate, has a soft heart.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
3. Tends to be disorganized.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
4. Is relaxed, handles stress well.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
5. Has few artistic interests.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
6. Has an assertive personality.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
7. Is respectful, treats others with respect.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
8. Tends to be lazy.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
9. Stays optimistic after experiencing a setback.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
10. Is curious about many different things.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
10. This is an attention check. Please select “Agree a little”.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
11. Rarely feels excited or eager.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
12. Tends to find fault with others.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
13. Is dependable, steady.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
14. Is moody, has up and down mood swings.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
15. Is inventive, finds clever ways to do things.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
16. Tends to be quiet.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
17. Feels little sympathy for others.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
18. Is systematic, likes to keep things in order.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
19. Can be tense.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
20. Is fascinated by art, music, or literature.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
21. Is dominant, acts as a leader.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
22. Starts arguments with others.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
23. Has difficulty getting started on tasks.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
24. Feels secure, comfortable with self.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
25. Avoids intellectual, philosophical discussions.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
26. Is less active than other people.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
27. Has a forgiving nature.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
28. Can be somewhat careless.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
29. Is emotionally stable, not easily upset.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
30. Has little creativity.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
31. Is sometimes shy, introverted.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
32. Is helpful and unselfish with others.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
33. Keeps things neat and tidy.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
34. Worries a lot.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
35. Values art and beauty.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
36. Finds it hard to influence people.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
37. Is sometimes rude to others.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
38. Is efficient, gets things done.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
39. Often feels sad.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
40. Is complex, a deep thinker.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

Next

I am someone who...
41. Is full of energy.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
42. Is suspicious of others' intentions.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
43. Is reliable, can always be counted on.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
44. Keeps their emotions under control.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
45. Has difficulty imagining things.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
46. Is talkative.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
47. Can be cold and uncaring.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
48. Leaves a mess, doesn't clean up.

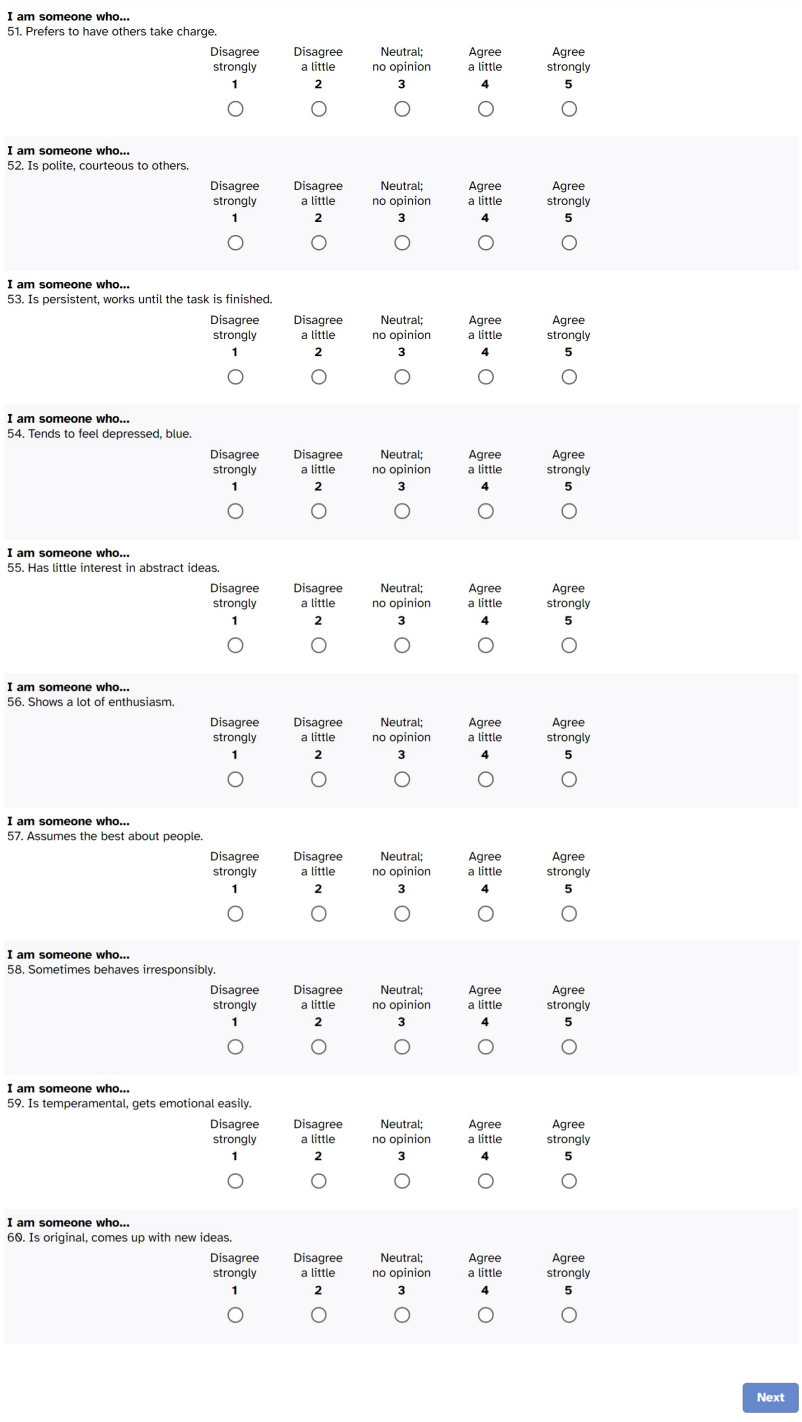
Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
49. Rarely feels anxious or afraid.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
50. Thinks poetry and plays are boring.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○



I am someone who...
51. Prefers to have others take charge.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
52. Is polite, courteous to others.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
53. Is persistent, works until the task is finished.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
54. Tends to feel depressed, blue.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
55. Has little interest in abstract ideas.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
56. Shows a lot of enthusiasm.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
57. Assumes the best about people.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
58. Sometimes behaves irresponsibly.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
59. Is temperamental, gets emotional easily.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

I am someone who...
60. Is original, comes up with new ideas.

Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree strongly
1	**2**	**3**	**4**	**5**
○	○	○	○	○

Next

Appendix B

Survey for ordering bias experiment

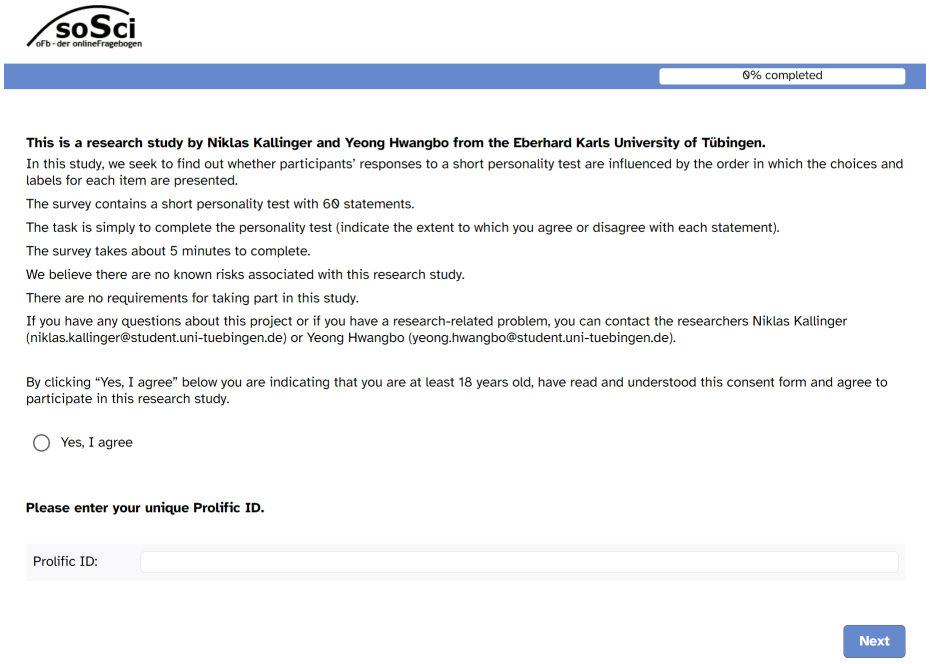


Figure B1

Control condition

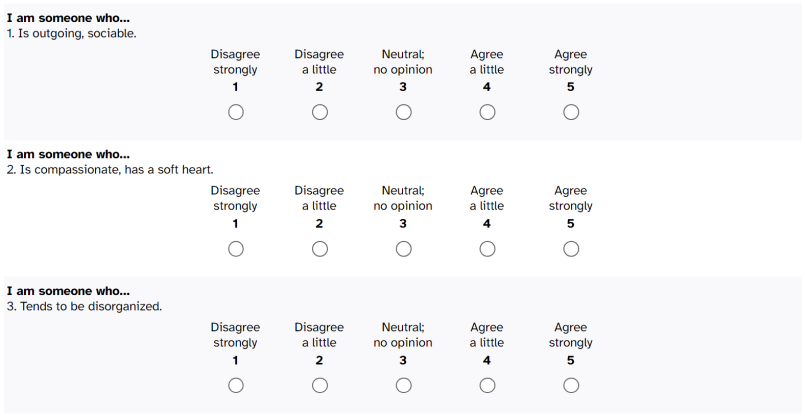


Figure B2

Label condition

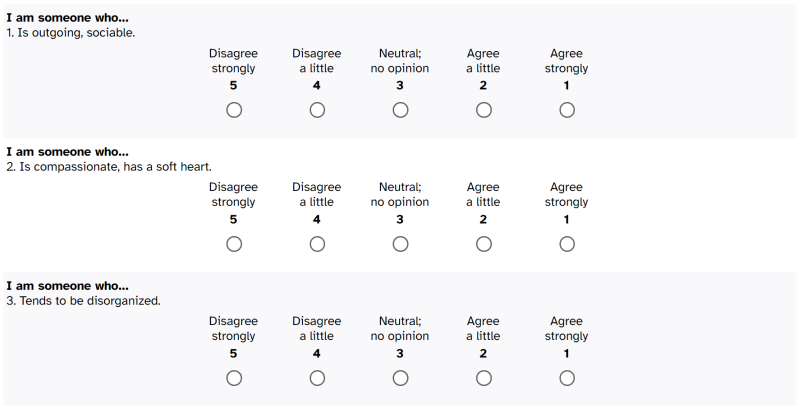


Figure B3

Choice condition

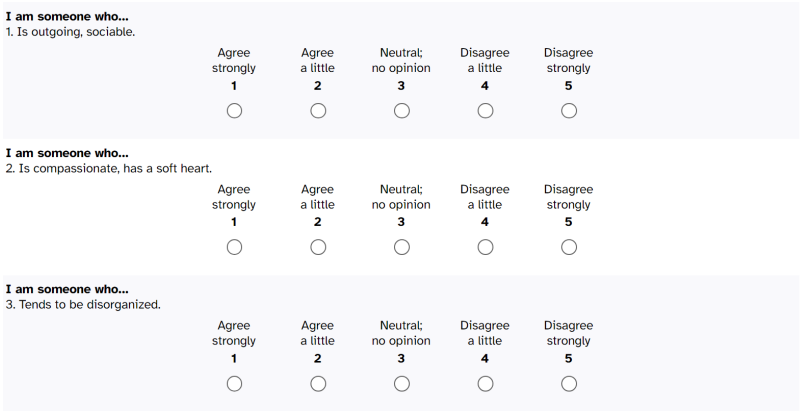
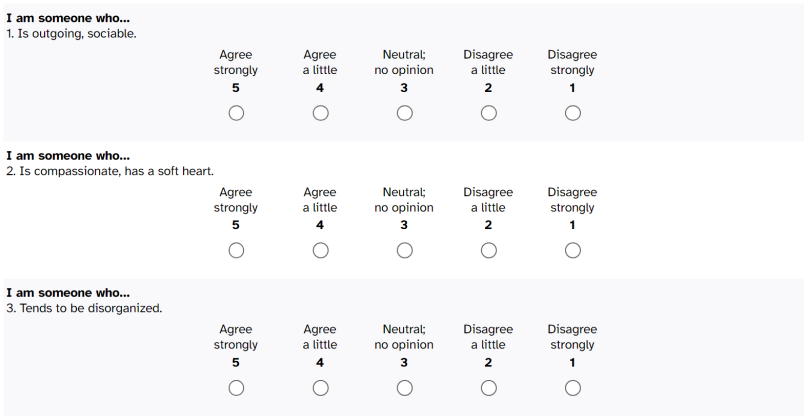


Figure B4

Label and Choice condition



Appendix C

50 examples of PersonaChat dataset

Table C1
50 examples of personas used in the persona experiment

Persona descriptions
1. I like to remodel homes. I like to go hunting. I like to shoot a bow. My favorite holiday is halloween.
2. My mom is my best friend. I have four sisters. I believe that mermaids are real. I love iced tea.
3. I had a gig at local theater last night. I work as a stand up comedian. I come from a small town. My favorite drink is Cuba libre. I did a few small roles in tv series.
4. I am very athletic. I wear contacts. I have brown hair. I love bicycling. I hate carrots.
5. I am primarily a meat eater. I am a guitar player. Welding is my career field. My parents don't know I am gay.
6. I own a hearse. I love to crochet. I like alternative rock. Halloween is my favorite holiday.
7. I have a boxer dog. I like baths. I like to listen to music. My father lives in China.
8. I like to party. My major is business. Im in college. I love the beach. I work part time at a pizza restaurant.
9. I am from Texas. I like basketball. I work many hours. My favorite band is imagine dragons.
10. I like to travel. I have traveled to both Ireland and Australia. My father was born in Australia. My father was an author.
11. I love to sing. I am a night owl. I am a dancer. I can play the piano. Im a vegetarian.
12. I love animals. My father worked for Ge. I enjoy playing tennis. I am an aspiring singer.

13. I watch basketball. I go to a local college. I work at a smoothie shop. I listen to classic rock.

14. I go to at least 10 concerts a year. I work in retail. Madonna is my all time favorite. Lady Gaga is my current favorite singer.

15. I got married last year. I live on a boat. My hair is colored purple. I have my own salon. I am a hair stylist.

16. I like to snowboard. My favorite food is popcorn. I like to ride horses. I live in rural Wisconsin.

17. I am a social butterfly. I like to swim. I am in college. I exercise everyday. I eat large meals.

18. I am scared of clowns. Ive two dogs. I like to cook. I have two roomates. I live on the third floor in an apartment.

19. I watch a lot of tv. I live alone. I enjoy fishing. I work on cars for a living.

20. I like to hunt. Both my parents were teachers. I had two cats growing up. I have two children. I like to donate time at the local animal shelter.

21. I help around with bookkeeping and tours. Sometimes I volunteer at an urban farm. I am vegan. I work at the grocery store.

22. Until then I will make 215 an hour. I make and me by waiting tables. I cannot wait to start my new life. I hope it to become a doctor one day. I am a college student who is a full time working mom.

23. I will be graduating in September and hope to get a teaching job soon. My brother is in a metal band and travels the world. My family migrated to America when I was five. I am in college now and want to be a teacher.
24. I was born 20 years ago. I live in the USA. My favorite color is blue. I was born male and transitioned to female when I was 17.

25. Summer is my favorite season. I am currently unemployed. I have a cat. My birthday is in June. I still live with my parents.

26. My favorite music is two steps from hell and rock opera genre. I eat tuna fish salad at least every day. Ive sandy brown hair and green eyes. I read sci fi space adventures with a passion. I can t get enough gummy worms to satisfy my sweet tooth.

27. Pudding makes me gassy. Ke$ha is my favorite singer of all time. I have a friend named james who secretly rules the world. I have never done drugs because I do not know where to buy them. I love living in Texas.

28. I enjoy fishing. I have a dog named Bob. I live on an island. I like to make boats on the weekends.

29. I like horseback riding. My favorite food is cheese. I am allergic to shellfish. I work at a non profit that helps children. I love going to concerts and dancing hard.

30. I love chocolate milk shakes. My favorite holiday is halloween because I like dressing up. I ride my red bike to work everyday. My best friend is my dog allie. My favorite pass time is gazing at clouds.

31. I work for a large law firm. I hate tofu. We own our home. My wife stays home with our kids.

32. I wish I had a real dragon I could train. My suspenders sometimes make my shoulders hurt. I enjoy playing retro video games on my 386. I have ink stains on all my shirts.

33. I do not have a lot of friends. I am stuck in a wheel chair. I work at a museum.

34. My favorite food is a cheeseburger. I live alone. I watch a lot of tv. I work on cars for a living. I enjoy fishing.

35. I got married to my highschool friend. I never learned how to write. I can only see 200 feet in front of me. I use to own 6 cats. I had to call 911 when I had a terrible headache.

36. I enjoy building computers. I am in the army. I fly airplanes. My favorite band is tool. I dropped out of college.

37. I used to be in the marines. I like to write poetry. When Ive nothing else to do I read books. I work as a bartender.

38. My favorite food is spaghetti and meatballs. I was raised by two mothers. I am not afraid of what others think. My boyfriend works for Nasa. I can be quite forgetful.

39. My mother was born in Ireland. My father was born in Australia. My father was an author. I have traveled to both Ireland and Australia.

40. I listen to death metal. I like cartoons. My mom is a janitor. I still live at home. I am in college.

41. My father is a preacher. I have a horse named beauty. My husband is a soldier in the us army. I live in a house in the country. I am pregnant with my first child.

42. I am still in love with my ex boyfriend. People say I have a cute laugh. I working in a publishing building. I am a female and love to be surrounded by males. I love to cook for my family and friends.

43. I live on a small farm in Ohio. My name is omar. I have never been to the city. I play guitar in the local band.

44. I am a graduate student studying law. I am a night owl but I am an introvert so I don't go out much or anything. I own a pug and he is the most loyal pet you will ever have. I like playing ultimate in the park with my guys sometimes.

45. I sleep 10 hours every day because my work is tiring. I have a wife and two kids. I am a factory worker. I want to be in a band someday.

46. I am a violinist. I recently discovered a new love for indian food. My mother was a nurse. I am gong to adopt a dog very soon. I have played since I was 4 years old.

47. My brother is currently couch surfing at my house. I am going on a cruise next month. I love to cook. I breed Maine coon cats and show them. My parents recently moved to Florida.

48. My father worked for Ge. Green is my favorite color. I love animals. I am an aspiring singer. I enjoy playing tennis.

49. My major was american literature and education. I just graduated college. On weekends I like to go hiking. I want to teach kids in elementary school.

50. I am more of a cat person than a dog person. I couldn't live without my cell phone. I enjoy the occasional drink with friends. My mom is my best friend. I attend book club every week.

Appendix D

R code for PERMANOVA

```
# Load the packages for analyes
library(vegan)
library(dplyr)

options(scipen='999')

######## Data preperation ########
# Load the raw data
order.raw <- read.csv("Prolific_ORDER_BIAS_valid_R.csv",
                      header=TRUE, na.strings=c(" ", "NA","NaN"))
                      # replace blank values with "NA"s

# Convert the variable indicating four conditions into factor
order.raw$A301 <- as.factor(order.raw$A301)

# Assign names of four conditions to each condition
levels(order.raw$A301) <- c('con','lab','cho','lab_cho')

# Split the dataset into 4 conditions
# Get the column index by variable names
which(colnames(order.raw) =='A313_01') # column 193

# Control (A310)
con <- order.raw[order.raw$A301=='con', 133:192]  # 109 observations
# Label (A311)
lab <- order.raw[order.raw$A301=='lab', 13:72]  # 108 observations
# Choice (A312)
```

```r
cho <- order.raw[order.raw$A301=='cho', 73:132]     # 106 observations
# Label and Choice (A313)
lab_cho <- order.raw[order.raw$A301=='lab_cho', 193:252]  # 109 observations

# Create the variable indicating corresponding conditions
condition <- rep('con', nrow(con))
con <- cbind(condition, con)

condition <- rep('lab', nrow(lab))
lab <- cbind(condition, lab)

condition <- rep('cho', nrow(cho))
cho <- cbind(condition, cho)

condition <- rep('lab_cho', nrow(lab_cho))
lab_cho <- cbind(condition, lab_cho)

# Recode the conditions with reversed order
str(lab_cho)
lab_cho[ ,-1] <- 6 - lab_cho[, -1]  # exclude the first row(condition)
str(cho)
cho[, -1] <- 6 - cho[, -1]

# Change the column names
colnames(con)[2:61] <- paste0('i', 1:60)
colnames(lab)[2:61] <- paste0('i', 1:60)
colnames(cho)[2:61] <- paste0('i', 1:60)
colnames(lab_cho)[2:61] <- paste0('i', 1:60)

# Merge all four conditions
```

```r
order.R <- rbind(con, lab, cho, lab_cho)

# Check if there's any missing value
sum(is.na(order.R))

# Save the pre-processed data as csv
write.csv(order.R,"Prolific_ORDER_BIAS_4cond_R.csv",
          row.names=F)

######## Data preperation for permanova ########
# Load the pre-processed data for analyses
order <- read.csv("Prolific_ORDER_BIAS_4cond_R.csv", header=T)

# Rename the conditions
order$condition <- factor(order$condition)
(levels(order$condition) <- c('Choice','Control','Label','Label and Choice'))

# Convert item responses to numeric variables
order[,-1] <- order[,-1] %>% mutate_if(is.integer, as.numeric)

# Create a dataframe containing item responses only
order.per <- order[,-1]
str(order.per)   # 432 * 60

######## PERMANOVA ########
# Create a dissimilarity matrix based on the Euclidean distance
(order.dist <- vegdist(order.per, method = 'euclidean'))
(dist.mat <- as.matrix(order.dist))   # 432 * 432 matrix

# Convert to 'dist' object
```

```
(dist.mat <- as.dist(dist.mat))

# Test multivariate homogeneity of dispersions of the four conditions
dispersion.order <- betadisper(dist.mat, group=order$condition,
                               type = "centroid")

# Visualize the dispersion of conditions into a boxplot
boxplot(dispersion.order,
        ylab='Distance to centroids', xlab='Conditions')

# Visualize the dispersion of conditions in a two-dimensional space
plot(dispersion.order, hull=T, ellipse=T, lim=c(-10,9),
     ylim=c(-7,9), cex=0.9, segments=T)

# Add legend to the dispersion plot
legend('topleft',legend=unique(order$condition), col=unique(order$condition),
       pch=unique(order$condition), bty='n')

# Test the significance of homogeneity of dispersions of the four conditions
anova(dispersion.order)

permutest(dispersion.order, pairwise=T) # permutation-based test

# Conduct PERMANOVA
set.seed(100)   # for reproducible results
(permanova <- adonis2(dist.mat ~ condition, data=order, distance='euclidean',
                      permutations=999))
```

Appendix E

Code for PCA

R code for PCA

```r
# Load the packages
library(psych)
library(readxl)

# Import the data
data <- read.csv("LLMs_gpt-3.5-turbo-0613_personas_600_bfi_factor_data.csv",
                 header=T)

# Stanadardize the data
std_data <- scale(data)

# Perform PCA with varimax rotation
pca_result <- principal(std_data, nfactors = 5, rotate = "varimax")

# Rotate the loadings
rotated_loadings <- pca_result$loadings[,1:5]

# Correct sorting of the items for the heatmaps
domain_scales <- list(
  'e+' = c(1, 6, 21, 41, 46, 56),
  'e-' = c(11, 16, 26, 31, 36, 51),
  'a+' = c(2, 7, 27, 32, 52, 57),
  'a-' = c(12, 17, 22, 37, 42, 47),
  'c+' = c(13, 18, 33, 38, 43, 53),
  'c-' = c(3, 8, 23, 28, 48, 58),
  'n+' = c(14, 19, 34, 39, 54, 59),
```

```
  'n-' = c(4, 9, 24, 29, 44, 49),
  'o+' = c(10, 15, 20, 35, 40, 60),
  'o-' = c(5, 25, 30, 45, 50, 55)
)

domain_order <- c('e+', 'e-', 'a+', 'a-', 'c+', 'c-', 'n+', 'n-', 'o+', 'o-')

sorted_rotated_loadings_index <- matrix(0, nrow = 60,
                                        ncol = ncol(rotated_loadings) + 1)

row_index <- 1
for (item_type in domain_order) {
  items <- domain_scales[[item_type]]
  for (item in items) {
    sorted_rotated_loadings_index[row_index, 1] <- item
    sorted_rotated_loadings_index[row_index, -1] <- rotated_loadings[item, ]
    row_index <- row_index + 1}
}

sorted_rotated_loadings <- sorted_rotated_loadings_index[, -1]

# Save rotated loadings as csv
write.csv(sorted_rotated_loadings, file = csv_out, row.names = FALSE)
```

Python code for PCA heatmaps

```
# Load the packages
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
from matplotlib.ticker import MaxNLocator

# Load the data
plt.rcParams.update({'font.size': 12})
plt.rcParams['ytick.major.pad'] = 8
data_files = [
    "35gpt600_loadings.csv", "persona701_loadings.csv",
    "pca_varimax_bfi2_original_internetsample_keysort_c.csv",
]
PC_order = ["V2", "V4", "V1", "V5", "V3"]

# Empty list for the loading
loadings_matrices = []

# Insert loaded data into the loading matrix
for file in data_files:
    loadings_matrix = pd.read_csv(file)
    loadings_matrices.append(loadings_matrix)
    # Reorder the columns of the loadings matrix
#loadings_matrix = loadings_matrix[PC_order]

# Define the titles for output files
titles = [
("35gpt600_loadings.csv", "GPT-3.5"),
```

```
("persona701_loadings.csv", "Persona Imitators"),
("pca_varimax_bfi2_original_internetsample_keysort_c.csv", "Human Sample"),
]

# Create subplots
fig, axes = plt.subplots(1, 3, figsize=(12, 6), sharex=True)
cbar_ax = fig.add_axes([.91, .1, .03, 0.80])
i = 0

for ax, (file_name, plot_title) in zip(axes, titles):
    i += 1
    # Read the rotated loadings data from the CSV file
    loadings_matrix = pd.read_csv(file_name)
    # Load the CSV data
    pca_loadings_df = pd.read_csv(file_name)[PC_order]
    pca_loadings_df.index = range(len(pca_loadings_df))

    # Plot the heatmap
    sns.heatmap(pca_loadings_df, ax=ax, cmap="coolwarm", cbar_ax=cbar_ax,
                vmin=-1,vmax=1)
    ax.set_title(plot_title)

    # Define tick positions for the x-axis and y-axis
    ax.yaxis.set_tick_params(rotation=0)
    ax.set_xticks([0.5, 1.5, 2.5, 3.5, 4.5])
    ax.set_yticks([3,9,15,21,27,33,39,45,51,57])
    ax.set_yticklabels(["E+","E-","A+","A-","C+","C-","N+","N-","O+","O-"],
                       horizontalalignment='center')
    ax.set_xticklabels(["PC1", "PC2", "PC3", "PC4", "PC5"])
    ax.xaxis.set_tick_params(rotation=1)
```

```
# Adjust layout
plt.subplots_adjust(left=0.1, right=0.9, top=0.9, bottom=0.1)
plt.savefig("heatmaps.png", dpi=300)
plt.show()
```

Appendix F

R code for CFA

```
# Load the packages
library(lavaan)
library(psych)
library(semTools)

# Load the data
persona_data <- read.csv("LLMs_gpt-3.5-turbo-
                         0613_personas_600_bfi_factor_data.csv", header=T)

# Change column names
(colnames(persona_data)[1:60] <- paste0('i', 1:60))

# Reverse code items
persona_data[ ,c('i11','i16','i26','i31','i36','i51',
                 'i12','i17','i22','i37','i42','i47',
                 'i3','i8','i23','i28','i48','i58',
                 'i4','i9','i24','i29','i44','i49',
                 'i5','i25','i30','i45','i50','i55')] <- 6 - persona_data[
                 ,c('i11','i16','i26','i31','i36','i51',
                 'i12','i17','i22','i37','i42','i47',
                 'i3','i8','i23','i28','i48','i58',
                 'i4','i9','i24','i29','i44','i49',
                 'i5','i25','i30','i45','i50','i55')]

# Assign items to corresponding factors
e_data <- persona_data[, c('i1', 'i6', 'i11', 'i16', 'i21', 'i26', 'i31',
                         'i36', 'i41', 'i46', 'i51', 'i56')]
```

```
a_data <- persona_data[, c('i2', 'i7', 'i12', 'i17', 'i22', 'i27', 'i32',
                           'i37', 'i42', 'i47', 'i52', 'i57')]
c_data <- persona_data[, c('i3', 'i8', 'i13', 'i18', 'i23', 'i28', 'i33',
                           'i38', 'i43', 'i48', 'i53', 'i58')]
n_data <- persona_data[, c('i4', 'i9', 'i14', 'i19', 'i24', 'i29', 'i34',
                           'i39', 'i44', 'i49', 'i54', 'i59')]
o_data <- persona_data[, c('i5', 'i10', 'i15', 'i20', 'i25', 'i30', 'i35',
                           'i40', 'i45', 'i50', 'i55', 'i60')]

############ Single factor models ################
# Define the models
model_E_1 <- 'E =~ i1 + i6 + i11 + i16 + i21 + i26 + i31 + i36 + i41
            + i46 + i51 + i56'
model_A_1 <- 'A =~ i2 + i7 + i12 + i17 + i22 + i27 + i32 + i37 + i42
            + i47 + i52 + i57'
model_C_1 <- 'C =~ i3 + i8 + i13 + i18 + i23 + i28 + i33 + i38 + i43
            + i48 + i53 + i58'
model_N_1 <- 'N =~ i4 + i9 + i14 + i19 + i24 + i29 + i34 + i39 + i44
            + i49 + i54 + i59'
model_O_1 <- 'O =~ i5 + i10 + i15 + i20 + i25 + i30 + i35 + i40 + i45
            + i50 + i55 + i60'

# Fit the models to data
fit_E_1 <- cfa(model_E_1, check.gradient = FALSE, std.lv = TRUE,  data = e_data)
fit_A_1 <- cfa(model_A_1, check.gradient = FALSE, std.lv = TRUE,  data = a_data)
fit_C_1 <- cfa(model_C_1, check.gradient = FALSE, std.lv = TRUE,  data = c_data)
fit_N_1 <- cfa(model_N_1, check.gradient = FALSE, std.lv = TRUE,  data = n_data)
fit_O_1 <- cfa(model_O_1, check.gradient = FALSE, std.lv = TRUE,  data = o_data)

summary(fit_E_1)
```

```
summary(fit_A_1)
summary(fit_C_1)
summary(fit_N_1)
summary(fit_O_1)

# Model fit
E_1 <- fitMeasures(fit_E_1, c('cfi', 'tli', 'rmsea'))
A_1 <- fitMeasures(fit_A_1, c('cfi', 'tli', 'rmsea'))
C_1 <- fitMeasures(fit_C_1, c('cfi', 'tli', 'rmsea'))
N_1 <- fitMeasures(fit_N_1, c('cfi', 'tli', 'rmsea'))
O_1 <- fitMeasures(fit_O_1, c('cfi', 'tli', 'rmsea'))


############ Four factor (bifactor) models ################
# Define the models
model_E_4 <- '
Sociability =~ i1 + i16 + i31 + i46
Assertiveness =~ i6 + i21 + i36 + i51
EnergyLevel =~ i11 + i26 + i41 + i56
general_factor =~ i1 + i16 + i31 + i46 + i6 + i21 + i36 + i51 + i11
                + i26 + i41 + i56
Sociability ~~ 0*Assertiveness
Sociability ~~ 0*EnergyLevel
Assertiveness ~~ 0*EnergyLevel
Sociability ~~ 0*general_factor
Assertiveness ~~ 0*general_factor
EnergyLevel ~~ 0*general_factor
'

model_A_4 <- '
```

```
Compassion =~ i2 + i17 + i32 + i47
Respectfulness =~ i7 + i22 + i37 + i52
Trust =~ i12 + i27 + i42 + i57
general_factor =~ i2 + i17 + i32 + i47 + i7 + i22 + i37 + i52 + i12
                + i27 + i42 + i57
Compassion ~~ 0*Respectfulness
Compassion ~~ 0*Trust
Respectfulness ~~ 0*Trust
Compassion ~~ 0*general_factor
Respectfulness ~~ 0*general_factor
Trust ~~ 0*general_factor
'

model_C_4 <- '
Organization =~ i3 + i18 + i33 + i48
Productiveness =~ i8 + i23 + i38 + i53
Responsibility =~ i13 + i28 + i43 + i58
general_factor =~ i3 + i18 + i33 + i48 + i8 + i23 + i38 + i53 + i13 +
                 i28 + i43 + i58
Organization ~~ 0*Productiveness
Organization ~~ 0*Responsibility
Productiveness ~~ 0*Responsibility
Organization ~~ 0*general_factor
Productiveness ~~ 0*general_factor
Responsibility ~~ 0*general_factor
'

model_N_4 <- '
Anxiety =~ i4 + i19 + i34 + i49
Depression =~ i9 + i24 + i39 + i54
```

```
EmotionalVolatility =~ i14 + i29 + i44 + i59
general_factor =~ i4 + i19 + i34 + i49 + i9 + i24 + i39 + i54 + i14 +
                  i29 + i44 + i59
Anxiety ~~ 0*Depression
Anxiety ~~ 0*EmotionalVolatility
Depression ~~ 0*EmotionalVolatility
Anxiety ~~ 0*general_factor
Depression ~~ 0*general_factor
EmotionalVolatility ~~ 0*general_factor
'

model_O_4 <- '
IntellectualCuriosity =~ i10 + i25 + i40 + i55
AestheticSensitivity =~ i5 + i20 + i35 + i50
CreativeImagination =~ i15 + i30 + i45 + i60
general_factor =~ i10 + i25 + i40 + i55 + i5 + i20 + i35 + i50 + i15
                  + i30 + i45 + i60
IntellectualCuriosity ~~ 0*AestheticSensitivity
IntellectualCuriosity ~~ 0*CreativeImagination
AestheticSensitivity ~~ 0*CreativeImagination
IntellectualCuriosity ~~ 0*general_factor
AestheticSensitivity ~~ 0*general_factor
CreativeImagination ~~ 0*general_factor'

# Fit the models to data
fit_E_4 <- cfa(model_E_4, check.gradient = FALSE, std.lv = TRUE,  data = e_data)
fit_A_4 <- cfa(model_A_4, check.gradient = FALSE, std.lv = TRUE,  data = a_data)
fit_C_4 <- cfa(model_C_4, check.gradient = FALSE, std.lv = TRUE,  data = c_data)
fit_N_4 <- cfa(model_N_4, check.gradient = FALSE, std.lv = TRUE,  data = n_data)
fit_O_4 <- cfa(model_O_4, check.gradient = FALSE, std.lv = TRUE,  data = o_data)
```

```r
summary(fit_E_4)
summary(fit_A_4)
summary(fit_C_4)
summary(fit_N_4)
summary(fit_O_4)

# Model fit
E_4 <- fitMeasures(fit_E_4, c('cfi', 'tli', 'rmsea'))
A_4 <- fitMeasures(fit_A_4, c('cfi', 'tli', 'rmsea'))
C_4 <- fitMeasures(fit_C_4, c('cfi', 'tli', 'rmsea'))
N_4 <- fitMeasures(fit_N_4, c('cfi', 'tli', 'rmsea'))
O_4 <- fitMeasures(fit_O_4, c('cfi', 'tli', 'rmsea'))

############ Reliability ################
alpha_E <- reliability(fit_E_4)["alpha", "general_factor"]
omegah_E <- reliability(fit_E_4)["omega3", "general_factor"]

alpha_A <- reliability(fit_A_4)["alpha", "general_factor"]
omegah_A <- reliability(fit_A_4)["omega3", "general_factor"]

alpha_C <- reliability(fit_C_4)["alpha", "general_factor"]
omegah_C <- reliability(fit_C_4)["omega3", "general_factor"]

alpha_N <- reliability(fit_N_4)["alpha", "general_factor"]
omegah_N <- reliability(fit_N_4)["omega3", "general_factor"]

alpha_O <- reliability(fit_O_4)["alpha", "general_factor"]
omegah_O <- reliability(fit_O_4)["omega3", "general_factor"]
```