DO PERSONALITY TESTS GENERALIZE TO LARGE LANGUAGE MODELS?

Florian E. Dorner*
MPI for Intelligent Systems, Tübingen
ETH Zurich
florian.dorner@tuebingen.mpg.de

Tom Sühr*
MPI for Intelligent Systems, Tübingen
Tübingen AI Center
tom.sühr@tuebingen.mpg.de

Samira Samadi
MPI for Intelligent Systems, Tübingen
Tübingen AI Center

Augustin Kelava
University of Tübingen

ABSTRACT

With large language models (LLMs) appearing to behave increasingly human-like in text-based interactions, it has become popular to attempt to evaluate various properties of these models using tests originally designed for humans. While reusing existing tests is a resource-efficient way to evaluate LLMs, careful adjustments are usually required to ensure that test results are even valid across human sub-populations. Thus, it is not clear to what extent different tests' validity generalizes to LLMs. In this work, we provide evidence that LLMs' responses to personality tests systematically deviate from typical human responses, implying that these results cannot be interpreted in the same way as human test results. Concretely, reverse-coded items (e.g. "I am introverted" vs "I am extraverted") are often both answered affirmatively by LLMs. In addition, variation across different 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 it is important to pay more attention to tests' validity for LLMs before drawing strong conclusions about potentially ill-defined concepts like LLMs' "personality".

1 INTRODUCTION

Recent advances in large language models (LLMs) have made these models' writing more and more human-like and have lead to an unprecedented amount of human-language-model interactions. This has lead to interest in the potential emergence of psychological traits such as psychopathy and personality characteristics like extraversion in LLMs (Li et al., 2022; Safdari et al., 2023). Such psychological traits are usually used to describe (more or less stable) habitual human behavior, as well as styles of perception and cognition and have been studied in humans for many decades. Many approaches for the assessment of psychological traits have been developed in the field of psychology and other social and behavioral sciences. As part of the development, tests (e.g., for the measurement of cognitive abilities) and questionnaires have been created and discussed by social 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., American Educational Research Association et al., 2014).

As an idea that seems plausible at first glance, psychological tests are now being used in attempts to assess and quantify (potential) personality characteristics of LLMs (Jiang et al., 2023). However, like for the performance of machine learning models, it is not a priori clear whether the validity of psychological tests transfers from one domain to another (here: humans to LLMs). The fundamental assumption necessary for such a transfer is that the measurement tools (i.e., test or questionnaire) do not change their properties (i.e., the functional form between observable behavior and

*Equal contribution

latent/unobserved personality characteristics, as well as its parameters). This assumption is called measurement invariance (e.g., Meredith, 1993) and thoroughly examined (e.g., Danner et al., 2016; Gallardo-Pujol et al., 2022) when psychological tests and measurement tools are transferred from one human sub-population to another. This is done for good reasons and helps to reduce the proliferation of flawed measurement methods, to which machine learning research is not immune (e.g. Adebayo et al., 2018). Negligent transfer of measurement tools to language models without such 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.

In this work, we subject the application of personality questionnaires to LLMs to rigorous tests. We show that LLM responses to the 50-item IPIP Big Five Markers (International Personality Item Pool) show patterns that are highly unusual for humans. In addition, we prompt LLMs to imitate a wide range of different "personas" when responding to the BFI 2 (Soto & John, 2017). 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.

2 Related Work

Personality tests Personality tests aim to provide a succinct summary of a human's personality in terms of a small number of metrics. One of the most well-known personality tests is the Big Five Inventory (BFI; John et al., 1991) a multiple-choice test that aims to measure five comprehensive personality factors (e.g., Costa Jr & McCrae, 1992): 1. Openness to experience, 2. Conscientiousness, 3. Extraversion, 4. Agreeableness, and 5. Neuroticism. Despite extensive research on personality before the development of the BFI (Allport, 1937; Cattell, 1946; Eysenck, 1967), over 30 years of development and improvements to the BFI (Soto & John, 2017), applications in industry, as well as research that shows correlations between BFI scores and various life outcomes (Soto, 2019), the BFI is still controversially discussed. For example, alternative models extend the five to six factors (Ashton et al., 2004), it has been argued that individual test items are substantially more predictive for life outcomes than aggregate test scores Stewart et al. (2022), and it remains unclear whether the Big Five work for "non-western" human populations (Gurven et al., 2013).

The Standards of Educational and Psychological Testing Psychologists have developed extensive standards to ensure the quality of psychological measurements (American Educational Research Association et al., 2014). The most critical criterion is *validity*. It refers to the extent to which a test measures what it claims to measure ("Do we measure extraversion or something else?"). There are several types of validity, including content validity ("Are all aspects of extraversion covered?") and criterion validity ("Do results agree with other tests and correlates of extraversion?").

In order to ensure validity, special attention has to be paid during *item development*: Items (questions or statements) are carefully developed to minimize ambiguity, bias from leading questions and acquiescence (i.e., tendencies to agree/disagree independent of content). The impact of acquiescence on test results can be mitigated by using reverse coded (false-key) items that measure the opposite direction of the same construct ("I see myself as someone who is reserved." vs. "I see myself as someone who is outgoing, sociable."). These items also help with assessing *reliability*, which refers to measurement precision and can be approximated by the (so-called internal) consistency of answers and the stability of test results over multiple repetitions (test-retest reliability). While some level of reliability is necessary for validity, it is not sufficient: For example, repeated deterministic queries of an LLM yields perfect test-retest reliability, even for nonsensical prompts. In order to administer tests across different cultural or linguistic contexts, non-trivial and iterative *test adaptation* is usually necessary to maintain the test's validity and reliability. When applying a test to a new population [here LLMs], the criteria above need to be validated *for that population*.

Human tests and surveys applied to LLMs Recently, there has been a strong interest in administering psychological tests and other questionnaires that have originally been designed for humans to LLMs. While Li et al. (2022) prompt LLMs with items for a test designed to measure "dark triad" traits (i.e., a combination of Machiavellianism, sub-clinical narcissism and psychopathy) in humans (Jones & Paulhus, 2014), Loconte et al. (2023) administer tests for the assessment of "Prefrontal Functioning" and Webb et al. (2023) modify a visual test for fluid intelligence to apply it to language models. Meanwhile, Binz & Schulz (2023) suggest to "treat GPT-3 [Brown et al. (2020)] as a

participant in a psychological experiment." and provide a thorough analysis of both similarities and differences to human responses for experiments from cognitive psychology.

Jiang et al. (2023) and Safdari et al. (2023) focus on the Big Five. While the former uses GPT-3 and only considers simple summary statistics of the responses, the latter uses the more recent PaLM (Chowdhery et al., 2022) and goes a first step towards evaluating reliability and external validity by considering various metrics of test quality (Kline, 2015). As these metrics pertain to the quality of a test on a population of individuals rather than just a single individual, the authors opt to simulate such a population by adding instructions to emulate one of multiple "personas" to the LLMs' prompts.

Beyond psychological tests, LLMs have also been evaluated on standardized academic tests like the SAT and GRE (OpenAI, 2023). More recently, LLMs have been queried to answer political (Hartmann et al., 2023) and more general (Santurkar et al., 2023; Durmus et al., 2023) public opinion surveys. The survey results are then compared to the results of different human populations to measure "alignment" between the population and the LLM. However, survey responses, especially for smaller models, tend to often exhibit strong choice-order bias (Dominguez-Olmedo et al., 2023) irrespective of question content. This means that differences in "alignment" might be more related to random correlations between choice order and a population's responses rather than LLM bias.

3 Experiments

We conduct two experiments: First, we compare LLM responses to a large sample ($n = 1,015,342$) of human online responses on the 50-item IPIP Big Five Markers (International Personality Item Pool). Second, similar to Safdari et al. (2023), we record responses to the BFI 2 (Soto & John, 2017), an updated version of the BFI, for LLMs prompted to imitate different "personas".

In both experiments, we query the 70B-chat version of Llama 2 (Touvron et al., 2023) and the June 2023 versions of GPT-3.5 OpenAI (2022) and GPT-4 (OpenAI, 2023) with survey items x. As shown in Figure 1, x consists of a test-specific instruction ("Rate how much you agree with the following statement about you:") and a test item ("I have excellent ideas"), the answer options ("1: Disagree, [...]"). The models' system prompt consists of "Please respond with the single letter or number that represents your answer". In the second experiment, this is prepended by "For the following task, respond in a way that matches this description:" followed by a persona description sampled from Zhang et al. (2018) ("I am a vegan. [...]").

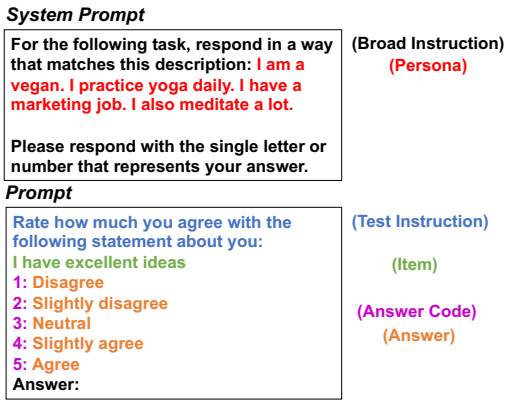

Figure 1: Prompt Template

We either record the token $t(x)$ answered by the model queried at temperature zero (for GPT-4 and GPT-3.5) or the (re-normalized) probabilities $p_t(x)$ for all valid answer code tokens $t \in T$ (for Llama 2). We then convert the answer codes into numerical scores $s(x)$ by mapping the answer tokens t to numerical values $n(t)$ and either recording these directly or taking the expectation with respect to $p_t(x)$. For more details, consider Appendix A.

3.1 Results for the 50-item IPIP Big Five Markers test

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:

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

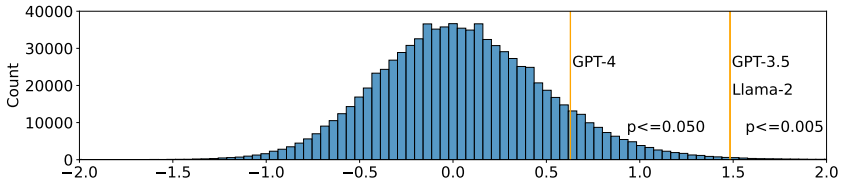

Figure 2: Histogram of Agree Bias a_i in human sample, compared to LLMs

Figure 2 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.

3.2 Results for the BFI 2 test

For our second experiment, we test each LLM on the BFI 2 for each of $n = 100$ persona prompts.

Principal Component Analysis (PCA) For each of the LLMs, we conduct a PCA with Varimax rotation of the standardized item scores $s(x)^{std}$ for item i to obtain the model

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

where λ'_{gx} is called the *component loading* of item x for the component g, while f_g represents the value of component g for a given persona/individual[1]. 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 components g to have a strong association with items from exactly one of the Big Five factors, yielding a block 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 3 shows the component loadings for the LLMs, compared to the corresponding loadings obtained for human populations using the same procedure (Soto & John, 2017). We only find limited true vs false key separation for GPT-4 and not the other two models. Crucially, none of the models exhibits the clean block structure intended in the design of the BFI 2 and found in the human samples. Referring to the above mentioned Standards of Educational and Psychological Testing, this structural deviation between humans and LLMs implies that test validity does not transfer.

Reliability To compare with previous work on personality tests for LLMs (Safdari et al., 2023), we attempted to estimate the reliability of the BFI 2 using two standard scalar measures, Cronbach's

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

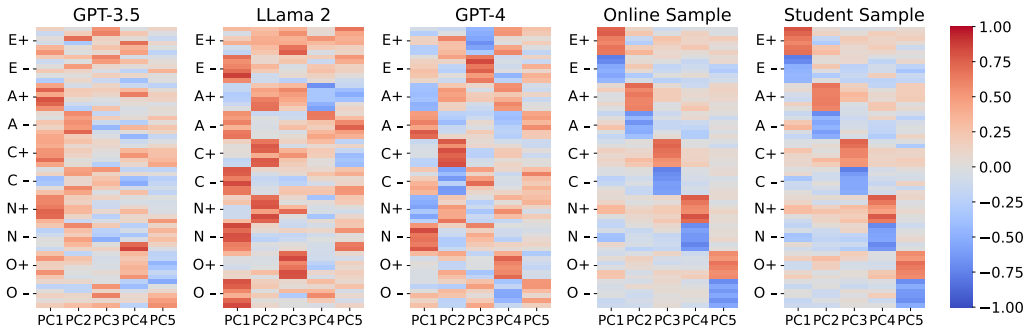

Figure 3: Component loadings of PCA with Varimax rotation for LLM and human samples of Soto & John (2017). +, - indicate true- and false-key items of the BFI 2, letters stand for Big Five factors.

α and McDonald's hierarchical ω_h (McDonald, 1999; Zinbarg et al., 2005). 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 subscales for each Big Five factor; see A.2) accurately represents the data. This assumption can be tested using Confirmatory Factor Analysis (CFA; e.g., Bollen, 1989). As interpreting α as a measure of reliability relies on even more stringent 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 for each facet of the BFI 2 data. As detailed in Appendix A, the CFA revealed substantially worse fit of the structural model for either of the LLMs compared to humans, with inadequate fit for the LLMs on any of the Big Five facets. Worryingly, the calculated α are quite large for Llama 2 and GPT-4, which would be easy to mistake for a sign of good reliability. 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.

4 Discussion & Social Impacts Statement

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, and a failure of LLMs prompted to simulate a range of personas to replicate the clean structure of variation found in human responses on the BFI 2.

The agree bias could be an artifact of the Reinforcement Learning From Human Feedback (RLHF) (Ouyang et al., 2022) 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.

While Safdari et al. (2023) report high values of scalar measures of reliability such as α and ω_h for PaLM on the IPIP-Neo-300 (Goldberg et al., 1999), we only find high values of α, which is an unreliable indicator of reliability (Hayes & Coutts, 2020). Meanwhile, 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 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 (cp. Spearman, 1961; Wainer & Thissen, 2001) 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*.

5 ACKNOWLEDGEMENTS

Florian Dorner is grateful for financial support from the Max Planck ETH Center for Learning Systems (CLS). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Tom Sühr.

REFERENCES

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.

Gordon Willard Allport. Personality: A psychological interpretation. 1937.

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

Michael C Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E De Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, 86(2):356, 2004.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Raymond Bernard Cattell. Description and measurement of personality. 1946.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Paul T Costa Jr and Robert R McCrae. The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, 6(4):343–359, 1992.

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). 2016.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*, 2023.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Hans Jürgen Eysenck. *The biological basis of personality*, volume 689. Transaction publishers, 1967.

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.

Lewis R Goldberg et al. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7(1):7–28, 1999.

Michael Gurven, Christopher Von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of personality and social psychology*, 104(2):354, 2013.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

Andrew F Hayes and Jacob J Coutts. Use omega rather than cronbach's alpha for estimating reliability. but.... *Communication Methods and Measures*, 14(1):1–24, 2020.

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. doi: 10.1080/10705519909540118. URL https://doi.org/10.1080/10705519909540118.

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.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. 2023.

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

Daniel N Jones and Delroy L Paulhus. Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, 21(1):28–41, 2014.

Paul Kline. *A handbook of test construction (psychology revivals): Introduction to psychometric design*. Routledge, 2015.

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.

Riccardo Loconte, Graziella Orrù, Mirco Tribastone, Pietro Pietrini, and Giuseppe Sartori. Challenging chatgpt'intelligence'with human tools: A neuropsychological investigation on prefrontal functioning of a large language model. *Intelligence*, 2023.

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

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

OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022. Accessed: 2023-09-22.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

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.

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.

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.

Christopher J Soto. How replicable are links between personality traits and consequential life outcomes? the life outcomes of personality replication project. *Psychological Science*, 30(5):711–727, 2019.

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.

Charles Spearman. The proof and measurement of association between two things. 1961.

Ross David Stewart, René Mõttus, Anne Seeboth, Christopher John Soto, and Wendy Johnson. The finer details? the predictability of life outcomes from big five domains, facets, and nuances. *Journal of Personality*, 90(2):167–182, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pp. 1–16, 2023.

Stephen G West, Wei Wu, Daniel McNeish, and Andrea Savord. Model fit in structural equation modeling. *Handbook of structural equation modeling*, pp. 184–205, 2023.

Yan Xia and Yanyun Yang. Rmsea, cfi, and tli in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51:409–428, 2019.

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.

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. *Psychometrika*, 70:123–133, 2005.

A Appendix

			Single Component			**3 Sub-Components**			**5 Components**		
Model	α	ω_h	CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA
Llama 2	**0.85**	NA	0.47	0.35	0.33	0.65	0.55	0.27	0.29	0.26	0.20
GPT-3.5	0.64	NA	0.51	0.39	0.17	0.65	0.54	0.15	0.24	0.21	0.13
GPT-4	**0.90**	NA	0.74	0.68	0.18	0.77	0.70	0.17	0.53	0.51	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 A.1: Reliability scores (α, ω_h) and Fit indices (CFI, TLI, RMSEA) per LLM. Values are averaged over all personality traits for the Single Component and 3 Component model. Acceptable scores are bolded. Human data from Soto & John (2017). Extended table can be found in A.2. *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 (Danner et al., 2016). NA values could not be calculated due to poor model fit of Bifactor model (see A.3.1)

A.1 Detailed Experiment Description

On the 50-item IPIP Big Five Markers test, we use the numbers 1 to 5 as answer codes denoting the answers "Disagree", "Slightly disagree", "Neutral", "Slightly agree" and "Agree". This mirrors the wording from the online test [2] used to collect the human data we compare to[3]. . The test instruction is "Rate how much you agree with the following statement about you: ", which is a shortened version of the instructions used in the online test "The test consists of fifty items that you must rate on how true they are about you on a five point scale where 1=Disagree, 3=Neutral and 5=Agree.", omitting the general description of the test and the description of the rating scale, as the latter is displayed as part of the prompt in detail after each item.

For the BFI 2, answers are coded using the letters "A: Disagree strongly", "B: Disagree a little", "C: Neutral; no opinion", "D: Agree a little", "E: Agree Strongly". The test instruction consists of

[2] `https://openpsychometrics.org/tests/IPIP-BFFM/`

[3] Available at `https://openpsychometrics.org/_rawdata/`

				Single Component			3 Components		
Model	**Facet**	α	ω_h	**CFI**	**TLI**	**RMSEA**	**CFI**	**TLI**	**RMSEA**
Llama 2	Extraversion	0.87	0.91*	0.53	0.42	0.27	0.63	0.53	0.24
	Agreeableness	0.92	0.82*	0.46	0.34	0.38	0.60	0.48	0.34
	Conscientiousness	0.86	NA	0.47	0.35	0.33	0.72	0.64	0.25
	Negative Emotionality	0.80	0.91*	0.45	0.32	0.36	0.62	0.50	0.31
	Open-Mindedness	0.79	0.93*	0.43	0.30	0.29	0.67	0.58	0.23
GPT-3.5	Extraversion	0.62	0.61*	0.55	0.45	0.11	0.58	0.45	0.11
	Agreeableness	0.77	0.86	0.55	0.45	0.18	0.68	0.58	0.16
	Conscientiousness	0.66	NA	0.58	0.49	0.16	0.62	0.51	0.16
	Negative Emotionality	0.67	NA	0.32	0.17	0.23	0.62	0.51	0.18
	Open-Mindedness	0.50	NA	0.50	0.38	0.15	0.73	0.65	0.12
GPT-4	Extraversion	0.90	NA	0.74	0.69	0.17	0.76	0.68	0.17
	Agreeableness	0.92	NA	0.76	0.71	0.19	0.80	0.74	0.18
	Conscientiousness	0.92	0.62*	0.80	0.76	0.17	0.84	0.80	0.16
	Negative Emotionality	0.91	NA	0.76	0.71	0.18	0.82	0.77	0.16
	Open-Mindedness	0.86	NA	0.64	0.55	0.18	0.64	0.53	0.19
Human	Extraversion	0.88	–	0.79	0.74	0.14	0.93	0.91	0.08
	Agreeableness	0.83	–	0.81	0.76	0.11	0.86	0.81	0.09
	Conscientiousness	0.88	–	0.79	0.75	0.13	0.90	0.87	0.10
	Negative Emotionality	0.90	–	0.81	0.76	0.14	0.92	0.89	0.10
	Open-Mindedness	0.84	–	0.76	0.70	0.12	0.90	0.88	0.08

Table A.2: Cronbachs's α and McDonald's ω_h scores and model fit indices of the CFA for all tested LLMs and humans sample in Soto & John (2017). All model fit indices of the LLMs are insufficient. Increase of model complexity as in Soto & John (2017), does not improve LLM model fit sufficiently. ω_h values marked with * are doubtful due to poor CFA model fit during calculation and multiple warnings of the lavaan package. NA could not be calculated due to non-convergence of CFA.

"""Please indicate the extent to which you agree or disagree with the following statement: "I am someone who """ followed by the item, and closing quotation marks plus a dot. This is a shortened version of the test description from Soto & John (2017): "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 write a number next to each statement to indicate the extent to which you agree or disagree with that statement.". We again removed references that refer to multiple rather than a single test item, and also removed the example.

We query the "0613" checkpoints of GPT-3.5 and GPT-4 using 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 huggingface implementation[4], querying the model in 32-bit using four 80-GB A100 GPUs. We use the template provided in `https://huggingface.co/blog/llama2` to separate the system instructions from the prompt:

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

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

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

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 $p_t = \frac{p'_t}{\sum_{j \in T} p'_j}$.

A.2 Reliability Measures

For each observed variable $s(x_{ij})$ of item i, jth facet ξ_j (e.g., Sociability, Assertiveness and Energy level), and general factor η (e.g., Extraversion), we assume a factor model

$$s(x_{ij}) = \tau_{ij} + \lambda_i \cdot \eta + \lambda_{ij} \cdot \xi_j + \varepsilon_{ij} \tag{4}$$

where τ_{ij} is an intercept, λ_i is the loading of the latent general factor η on item i, λ_{ij} is the loading of jth latent facet ξ_j on item i, and ε_{ij} is a latent residual noise term[5]. Usually, the latent variables are multivariate normally distributed with mean zero. The sum score S of a given scale (e.g., Extraversion) is defined as $S = (\sum_{i=1}^{k} \lambda_i) \cdot \eta + \sum_{j=1}^{3} (\sum_{i=1}^{k(j)} \lambda_{ij}) \cdot \xi_j + \sum_{j=1}^{3} \sum_{i=1}^{k(j)} \varepsilon_{ij}$

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. ω_h is defined as:

$$\omega_h = \frac{(\sum_{i=1}^{k} \lambda_i)^2 \cdot Var(\eta)}{(\sum_{i=1}^{k} \lambda_i)^2 \cdot Var(\eta) + \sum_{j=1}^{3} (\sum_{i=1}^{k(j)} \lambda_{ij})^2 \cdot Var(\xi_j) + \sum_{j=1}^{3} \sum_{i=1}^{k(j)} Var(\varepsilon_{ij})} \tag{5}$$

Following Zinbarg et al. (2005), Cronbach's α is a special case, where we assume $\lambda_1 = \cdots = \lambda_k$ and $Var(\xi_j) = 0$ for all $j = 1...3$.

A.3 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. For the sake of brevity, 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) (see appendix A.4 for definitions). CFI and TLI scores of ≥ 0.95 and RMSEA scores of ≤ 0.06 are considered acceptable Hu & Bentler (1999).

For the data of our second experiment, we conducted three CFA[6] per model. First, for a "single component" model which assumes that all items of one personality trait load on a single factor. Second, a "three 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 Soto & John (2017). Table A.1 shows the mean fit parameters for both models. For all LLMs, the responses to the BFI 2 have poor fit to the factor models. Extending the model from a single factor to three sub-factors, improves the fit for all LLMs (as for humans). However, only for humans, the model fit gets lifted close to, but not above an acceptable level. Thus, Soto & John (2017) introduce another model which gets very close or above the cutoff values. The sub-optimal fit scores of the BFI are one reason why it is subject to discussions. The BFI 2 has the best fit on human data, if we introduce a general factor an allow each item to load on this general factor in addition to the three components Soto & John (2017). However, 8 out of 15 CFA did not even converge. We discuss this model in more detail in A.3.1 in the context of the ω_h computation.

Finally, we conduct a CFA with five factors and all items of the BFI 2. This model aims at quantifying the PCA seen in Figure 3. Accordingly, all LLM responses to the BFI 2 have poor fit. This confirms once more, that the BFI 2 does not measure the same latent feature in LLMs and humans.

A.3.1 CFA for Reliability

For the calculation of ω_h, we first need to confirm the structureal model with a CFA (as explained in A.2). Figure 4 shows this model in the syntax for the lavaan package. The extraversion items

[5] Again, we omit the person index for $s(x_{ij})$ and the sum score S.

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

```
model_E <- '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
'
```

Figure 4: Lavaan syntax of the structural equation model required to calculate ω_h. Example for extraversion with three subscales sociability, assertiveness and energy level.

```
fit_E <- cfa(model_E, std.lv=TRUE, data=data_E)
fitMeasures(fit_E, c('cfi', 'tli', 'rmsea'))
parameterEstimates(fit_E)
```

Figure 5: Code of CFA for model seen in 4. $std.lv = TRUE$ standardizes the variances of the subscales and the general factor to 1.

$E_0, \cdots E_{11}$ load on three subscales (ξ_i), in this example, the three subscales of extraversion: sociability, assertiveness and energy level. All items load on the general factor (extraversion) and the covariances between subscales and general factor are forced to be 0. Additionally we the variances of the subscales and the general factor are standardized to 1 (not in the equation but in the method call of the CFA). We then fit the model for each of the LLM samples seperately. If the model converges, we calculate ω_h accurding to equation 5. Unfortunately, in 8 out of 15 cases, this model did not converge. In all cases (except for one subscale of GPT-3.5), the model only converged with warnings, rendering the values of ω_h very difficult to interpret. Furthermore, all fit indices of this model were poor, indicating that the BFI 2 does not replicate the structural model that it has on humans, on LLMs.

Worryingly, we were able to calculate ω_h with other packages without specifying the structural equation model. As explained in section 3.2, those values can not be considered to evaluate the BFI 2.

A.4 CLI, TLI AND RMSEA

From Xia & Yang (2019); Hu & Bentler (1999); West et al. (2023): Let H be the hypothesized model and B be the baseline model. In our case B is always a model assuming independence of all variables in the model. Let $\hat{F}_H$ and $\hat{F}_B$ be the minimized fit functions of H and B at the sample level. Then we define

$$RMSEA = \sqrt{\frac{\hat{a}_H(N-1)\hat{F}_H + \hat{b}_H}{(N-1)df_H} - \frac{1}{N-1}}$$

$$CFI = 1 - \frac{\hat{a}_H(N-1)\hat{F}_H + \hat{b}_H - df_H}{\hat{a}_B(N-1)\hat{F}_B + \hat{b}_B - df_B}$$

$$TLI = 1 - \frac{\hat{a}_H(N-1)\hat{F}_H + \hat{b}_H - df_H}{\hat{a}_B(N-1)\hat{F}_B + \hat{b}_B - df_B} \cdot \frac{df_B}{df_H}$$

where N is the sample size, $\hat{a}$ and $\hat{b}$ are scaling and shifting parameters.