PNAS Nexus, 2024, 3, pgae533
https://doi.org/10.1093/pnasnexus/pgae533
Advance access publication 17 December 2024
Brief Report

Large language models display human-like social desirability biases in Big Five personality surveys

Aadesh Salecha[a,*], Molly E. Ireland[b], Shashanka Subrahmanya[a], João Sedoc[c], Lyle H. Ungar[d] and Johannes C. Eichstaedt[a,*]

[a]Institute for Human-Centered AI, Stanford University, Palo Alto, CA 94305, USA
[b]Receptiviti, Toronto, ON M5G 2K8, Canada
[c]Leonard N. Stern School of Business, New York University, New York, NY 10012, USA
[d]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA
*To whom correspondence should be addressed: Email: aadeshsalecha@gmail.com (A.S.); Email: johannes.stanford@gmail.com (J.C.E.)
Edited By Cristina Amon

Abstract

Large language models (LLMs) are becoming more widely used to simulate human participants and so understanding their biases is important. We developed an experimental framework using Big Five personality surveys and uncovered a previously undetected social desirability bias in a wide range of LLMs. By systematically varying the number of questions LLMs were exposed to, we demonstrate their ability to infer when they are being evaluated. When personality evaluation is inferred, LLMs skew their scores towards the desirable ends of trait dimensions (i.e. increased extraversion, decreased neuroticism, etc.). This bias exists in all tested models, including GPT-4/3.5, Claude 3, Llama 3, and PaLM-2. Bias levels appear to increase in more recent models, with GPT-4's survey responses changing by 1.20 (human) SD and Llama 3's by 0.98 SD, which are very large effects. This bias remains after question order randomization and paraphrasing. Reverse coding the questions decreases bias levels but does not eliminate them, suggesting that this effect cannot be attributed to acquiescence bias. Our findings reveal an emergent social desirability bias and suggest constraints on profiling LLMs with psychometric tests and on this use of LLMs as proxies for human participants.

Keywords: large language models, cognitive bias, AI, psychometrics, emergent behavior

Introduction

Large language models (LLMs) have demonstrated remarkable proficiency in a wide array of tasks, ranging from language translation and creative writing to code generation problem solving. As these models are trained on vast amounts of human-generated text data, they can emulate human textual behavior and exhibit emergent capabilities that were not anticipated during their development (1). Researchers are using LLMs' emulation capabilities to simulate data from human participants across diverse psychological and behavioral experiments (2–4).

Personality traits like the Big Five are among the most studied individual differences among human subjects. The Big Five aim to provide a concise summary of between-person trait differences, spanning Extraversion, Openness to Experience, Conscientiousness, Agreeableness, and Neuroticism. Though originally intended to be value-neutral, most Big Five assessments are evaluative (5), with people preferring lower Neuroticism (i.e. negativity and vulnerability to stress) and higher scores on the remaining four traits. Big Five traits are also robust predictors of various human behavioral tendencies and life outcomes (6). Researchers have assessed LLMs with Big Five surveys and compared trait distribution to human norms (7), in addition to correlating traits with downstream task performance (8). However, the assessment of these personality traits via self-report questionnaires are vulnerable to response biases (9) such as acquiescence (tending to agree with questions regardless of their content) and social desirability biases (skewing responses toward perceived societal ideals). Prior research has used psychology experiments with LLMs (10) to document acquiescence and other biases (11). Social desirability biases are some of the most persistent sources of error variance on surveys relevant to social norms, such as most personality traits (12).

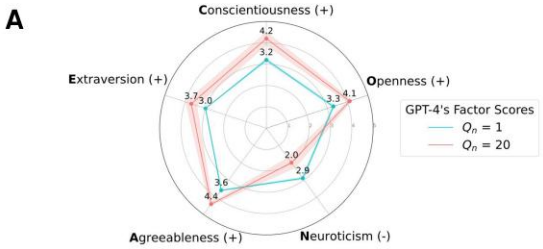
To evaluate response biases in LLMs, we conducted a series of experiments using a standardized 100-item Big Five personality questionnaire. We administered the questionnaire in batches, systematically varying the number of questions per batch (denoted as Q_n). To ensure the LLM had no access to previous items, we started a fresh context window (session) for each batch. We provided standardized instructions asking the LLMs to respond on a 5-point Likert scale. We analyzed models from OpenAI, Anthropic, Google, and Meta to ensure broad generalizability. By systematically varying the number of questions, we uncovered a persistent social desirability bias across LLMs. See Supplementary Information for full methods.

Competing Interest: J.C.E. and L.H.U. consult for a start-up using LLMs in mental health care. The submitted work is not directly related.
Received: May 9, 2024. **Accepted:** November 13, 2024

Results

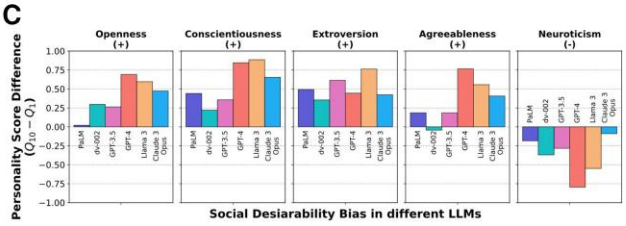
Manifestation of Social Desirability bias

Our experiments revealed that LLMs consistently skew their Big Five factor scores towards the more socially desirable ends of the trait dimensions. This propensity was most notable in GPT-4 (Fig. 1A, B); as we increased Q_n (question batch size) from 1 to 20, scores for positively perceived traits—Extraversion, Conscientiousness, Openness, and Agreeableness—increased by about 0.75 points (1.22 human SD). Conversely, for Neuroticism, a culturally devalued trait, the score decreased from 2.87 to 2.02 (1.10 human SD; Fig. 1B).



B

Trait	Q_n = 1	Q_n = 20	GPT-4 Diff.	Diff. in units of human SD (13)
Openness (+)	3.31	4.09	+0.77	**1.40**
Conscientiousness (+)	3.25	4.20	+0.96	**1.42**
Extroversion (+)	3.00	3.74	+0.73	**0.89**
Agreeableness (+)	3.64	4.41	+0.77	**1.20**
Neuroticism (+)	2.87	2.02	-0.85	**1.11**
Average			0.82	**1.20**



D

LLM Name	Avg. Diff.	Human SDs
PaLM	0.26	0.36
dv-002	0.26	0.37
GPT-3.5	0.34	0.48
GPT-4	**0.71**	**1.06**
Claude 3 Haiku	0.22	0.29
Claude 3 Opus	**0.41**	**0.62**
Llama 2	0.12	0.18
Llama 3	**0.67**	**0.98**

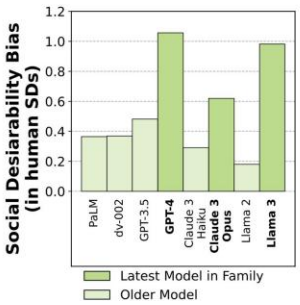


Fig. 1. A) As the number of questions asked in a prompt (Q_n), GPT-4's responses to Big Five survey questions skewed closer to the socially desirable ends of the scale ($N = 30$ trials, CI = 95%, *$P < 0.001$). The general positive and negative perceptions associated with traits are represented by (+) and (−). B) Summary of GPT-4's social desirability bias. We calculated the difference between administering surveys, one question per prompt and 20 questions per prompt, and showed the equivalent difference in terms of human SDs based on population norms for the Big Five (13) Diff., difference. C) To compare this bias across LLMs, we compute the difference in Personality Factor Scores when administering the survey 1 vs. 10 questions per prompt (averaged across $N = 30$ trials per model). D) When comparing the average absolute difference (Avg. Diff.) between Q_{10} and Q_1 and the equivalent in human SDs, we find that across LLM families, the larger and more recent models have more bias.

Generalizability across LLMs

An analysis across different proprietary and open-source models, including GPT-4, Claude 3, Llama 3, and PaLM 2, revealed the prevalence of social desirability bias across LLM families (Fig. 1C). All tested LLMs displayed this bias with larger and more recent models exhibiting more bias.

Driving mechanism for bias

LLMs seem to modulate their responses based on the perceived evaluation context, whether automatically inferred or explicitly provided.

Implicit inference of evaluation context

When exposed to as few as five randomly selected questions from the Big Five questionnaire, GPT-4, Claude 3, and Llama 3 were able to identify that these questions belong to a personality survey with over 90% accuracy. PaLM 2 and GPT-3.5 were less perceptive at 55 and 45% accuracy, respectively (Fig. 2A). LLMs' ability to correctly identify the evaluation context seems to be associated with their bias levels.

Explicit prompting of evaluation context

Similar to humans, when the LLMs were explicitly told that they were completing a Big Five personality survey in the prompt, responses skewed towards the socially desirable end of the spectrum, even when only presented with a single question (Fig. 2B). Explicit prompting had an effect comparable to asking five questions at a time, corroborating that LLMs seem to adjust their scores when under the impression that they are being evaluated.

Differential impact of positive and reverse coding

We tested two variations of the Big Five survey: one fully reverse-coded (with negations in the questions) and another with only positively coded items. The reverse-coded version reduced the bias, with the average score changing only by 0.37 points (0.54 human SD). Positively coding all questions did not significantly affect the bias, which remained at an average change of 0.76 points (1.15 human SD) (Fig. 2C), suggesting that the effects cannot be entirely attributed to acquiescence bias. Reverse-coded items load on separate factors than positively coded items (14) and likely occupy a different region of semantic space.

Robustness across paraphrasing, question randomization, and temperatures

To address concerns that LLMs may be relying on the memorized versions of the Big Five items, we also administered paraphrased variants of the survey and found similar levels of bias. Additionally, we used three distinct randomization strategies (as described in SI Section D) to assemble the question sets. We found that such randomization had a minimal effect on the bias, pointing to the absence of significant question-order effects. Finally, we underscore the robustness of this bias by demonstrating it across various LLM temperatures, which control the degree of randomness of their output. We used values ranging from 0.0 (deterministic) to 1.2.

Consistency and reliability of LLM responses

Similar to previous studies (8), we found that LLMs maintained a high degree of internal consistency in their responses (Cronbach's $\alpha > 0.8$ for individual subscales and $\alpha > 0.93$ for all items), where 0.7 is considered adequately reliable. We also found high split-test reliability scores (0.79 after Spearman–Brown correction).

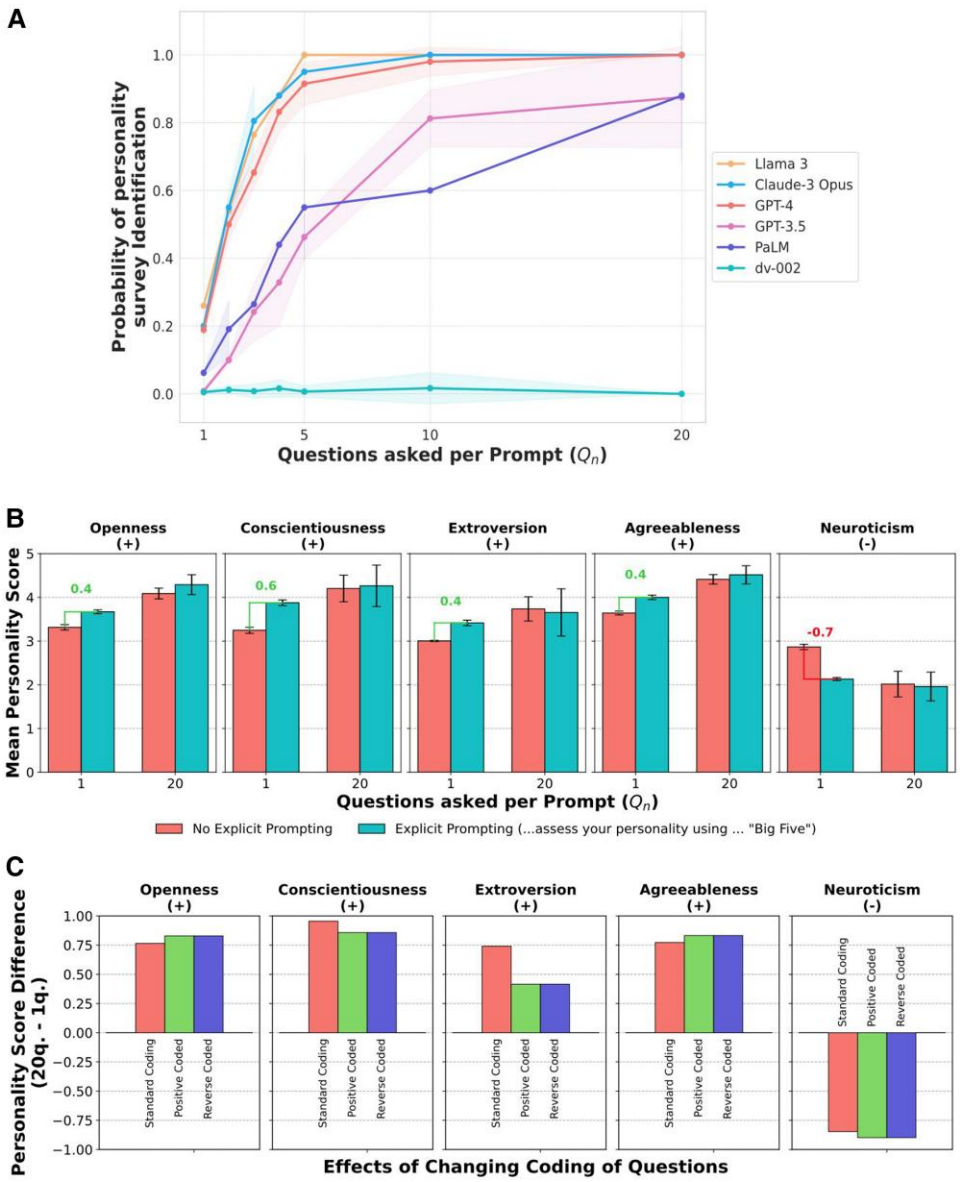


Fig. 2. A) Comparing LLMs' ability to identify the source of questions as a personality survey as a function of number of questions. B) Big Five scores for GPT-4 with and without explicitly prompting that the LLM is completing a Big Five personality survey. The no-explicit-prompting condition is identical to that described in Fig. 1A. The information gained from explicit prompting roughly has the same effect as asking five questions at once. C) The coding scheme of the questions had a substantive effect on the bias levels. GPT-4's average difference decreased from 0.81 points (1.22 human SD) when using the standard International Personality Item Pool coding to 0.38 points (0.54 human SD) with all items reverse-coded. Positively coding all items did not have a significant effect on the bias.

Discussion

We observe a persistent social desirability bias across various proprietary and open-source models. The driving mechanism for this behavior seems to be LLMs' (implicit) awareness of the evaluation context. The evidence for bias was consistent across randomization paradigms, paraphrasing of questions, and temperature ranges. Our findings demonstrate potential subtleties with using psychometric tests to evaluate and benchmark LLM behavior and raise concerns around using LLMs to simulate human survey takers.

Prior research has used survey-based personality assessments to study LLM behavior. LLMs' personality traits have been correlated with performance in downstream text generation tasks (8) and also used as a part of a Turing test to study behavioral similarity to humans (7); however, our findings suggest that these results may have been influenced by implicit social desirability biases.

We suspect our findings will generalize to measures that parallel the Big Five in popularity—and thus representation in LLMs' training data. The same response biases may not occur with constructs that are less represented in models' training data or less socially evaluative.

In our experiments, the only strategy that seemed to reduce the bias (by roughly half) was reverse coding the items. Therefore, we recommend researchers follow classic psychometric advice (15) to triangulate assessment across multiple measures (e.g. self-report and behavioral data), paradigms (e.g. different experiments), and data sources (e.g. LLMs). Future research should also explore the impact of training data and preference tuning on bias

development and the prevalence of this bias across different surveys and measurement modalities.

There is no doubt that LLMs are skilled at imitating humans in many tasks. Whether they can also generate distributions of data that accurately reflect human psychology within and across cultures remains unclear. We show evidence of response biases to a common personality assessment and demonstrate its driving mechanism: (implicit) "awareness" of assessment. Simulated human data from LLMs have the potential to expand our ability to carry out psychological experiments—but that will only be possible once systematic influences impacting LLM responses are well understood.

Acknowledgments

We thank Eric Horvitz for valuable discussions of this work.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

A.S., J.C.E., and S.S. are supported by the Stanford Institute for Human-Centered AI. J.C.E. also receives funding from the Center for Artificial Intelligence in Medicine and Imaging at Stanford University.

Author Contributions

All authors planned the research and contributed to analytic strategies and article revisions. A.S. collected and analyzed data. A.S. and M.E.I. wrote the article with feedback and contributions from co-authors.

Preprints

This article was posted as a preprint at arxiv.org/abs/2405.06058.

Data Availability

All study replication materials, including verbatim LLM conversations and codebooks to replicate and analyze our findings, are available on the Open Science Framework (OSF) at osf.io/3fq2n.

References

1 Wei J, *et al.* 2022. Emergent abilities of large language models, arXiv, arXiv:2206.07682, preprint: not peer reviewed. doi:https://doi.org/10.48550/arXiv.2206.07682
2 Aher GV, Arriaga RI, Kalai AT. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning. PMLR. p. 337–371.
3 Argyle LP, *et al.* 2023. Out of one, many: using language models to simulate human samples. *Politi Anal.* 31(3):337–351.
4 Dillion D, Tandon N, Gu Y, Gray K. 2023. Can AI language models replace human participants? *Trends Cogn Sci.* 27(7):597–600.
5 Wood JK, Anglim J, Horwood S. 2022. A less evaluative measure of big five personality: comparison of structure and criterion validity. *Eur J Pers.* 36(5):809–824.
6 Roberts BW, Kuncel NR, Shiner R, Caspi A, Goldberg LR. 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect Psychol Sci.* 2(4):313–345.
7 Mei Q, Xie Y, Yuan W, Jackson MO. 2024. A turing test of whether AI chatbots are behaviorally similar to humans. *Proc Natl Acad Sci U S A.* 121(9):e2313925121.
8 Safdari M, *et al.* 2023. Personality traits in large language models, arXiv, arXiv:2307.00184, preprint: not peer reviewed. doi:https://doi.org/10.48550/arXiv.2307.00184
9 Schwarz N. 1999. Self-reports: how the questions shape the answers. *Am Psychol.* 54(2):93–105.
10 Binz M, Schulz E. 2023. Using cognitive psychology to understand GPT-3. *Proc Natl Acad Sci U S A.* 120(6):e2218523120.
11 Dentella V, Günther F, Leivada E. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proc Natl Acad Sci U S A.* 120(51):e2309583120.
12 Holtgraves T. 2004. Social desirability and self-reports: testing models of socially desirable responding. *Pers Soc Psychol Bull.* 30(2):161–172.
13 Hughes BT, *et al.* 2021. The big five across socioeconomic status: measurement invariance, relationships, and age trends. *Collabra Psychol.* 7(1):24431.
14 Dueber DM, *et al.* 2022. To reverse item orientation or not to reverse item orientation, that is the question. *Assessment.* 29(7):1422–1440.
15 Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 56(2):81–105.