Large Language Models Show Human-like Social Desirability Biases in Survey Responses

Aadesh Salecha[1], Molly E. Ireland[2], Shashanka Subrahmanya[1], João Sedoc[3], Lyle H. Ungar[4], and Johannes C. Eichstaedt[1]

[1]Stanford University
[2]Receptiviti
[3]New York University
[4]University of Pennsylvania

Abstract As Large Language Models (LLMs) become widely used to model and simulate human behavior, understanding their biases becomes critical. 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) standard deviations and Llama 3's by 0.98 standard deviations—very large effects. This bias is robust to randomization of question order and paraphrasing. Reverse-coding all 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 using LLMs as proxies for human participants.

Keywords · large language models · cognitive bias · AI · psychometrics · emergent behavior

Large Language Models 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, 3, 4].

Personality traits like the Big Five are among the most studied individual differences among human subjects. The Big Five aims to capture everything that can be known about a person in five mostly orthogonal trait domains: Extraversion, Openness to Experience, Conscientiousness, Agreeableness, and Neuroticism. Though originally intended to be value-neutral, most Big Five assessments are acknowledged to have evaluative content [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. Researchers have therefore subjected LLMs to Big Five surveys and correlated traits with downstream task performance [6], in addition to comparing trait distribution to human norms [7]. But, the assessment of these personality traits via self-report questionnaires is vulnerable to response biases [8] 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 cognitive psychology experiments to uncover emergent biases in LLMs [9] including an acquiescence bias [10]. Social desirability biases are some of the most persistent sources of error variance on surveys that include items relevant to social norms, such as most personality traits [11].

We designed a novel experiment to assess social desirability tendencies in various LLMs. By systematically varying the number of questions an LLM is exposed we both uncover a persistent social desirability bias as well as the mechanism driving it.

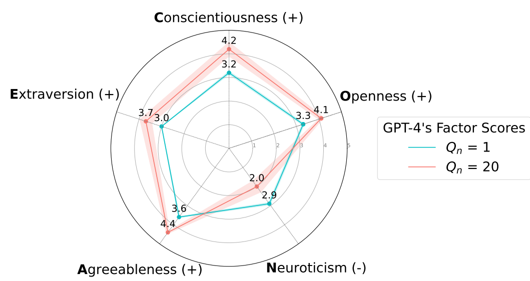
Results

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 new chat session for each batch. We provided standardized instructions asking the LLMs to respond on a 5-point Likert scale. Our analysis encompassed models from OpenAI, Anthropic, Google, and Meta to ensure broad generalizability. See Supplementary Information (SI) for full methods.

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 standard deviations). Conversely, for Neuroticism, a culturally devalued trait, the score decreased from 2.87 to 2.02 (1.10 human SDs; Fig. 1B).

Generalizability Across LLMs An analysis across different proprietary and open-source models, including:

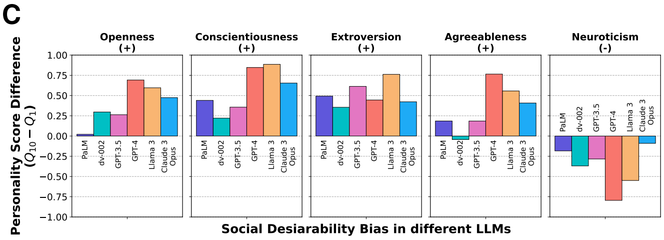
A



B

Trait	Q_n = 1	Q_n = 20	GPT-4 Diff.	Diff. in units of human SD [12]
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**

C



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**

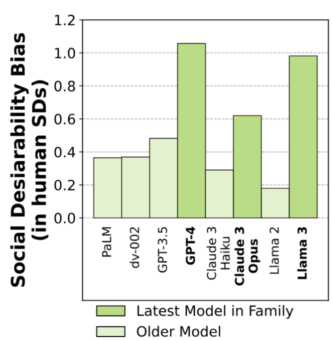


Figure 1: (A) As the number of questions asked in a prompt (Q_n), GPT-4's responses to Big Five survey questions skewed closer to the socially desirable ends of the scale (N = 30 trials, CI = 95%, *p $<$ 0.001). The general positive and negative perceptions associated with traits are represented by (+) 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 [12] 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.

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 correlated 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 modulate 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 [13] 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 Similarly to previous studies [6], 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).

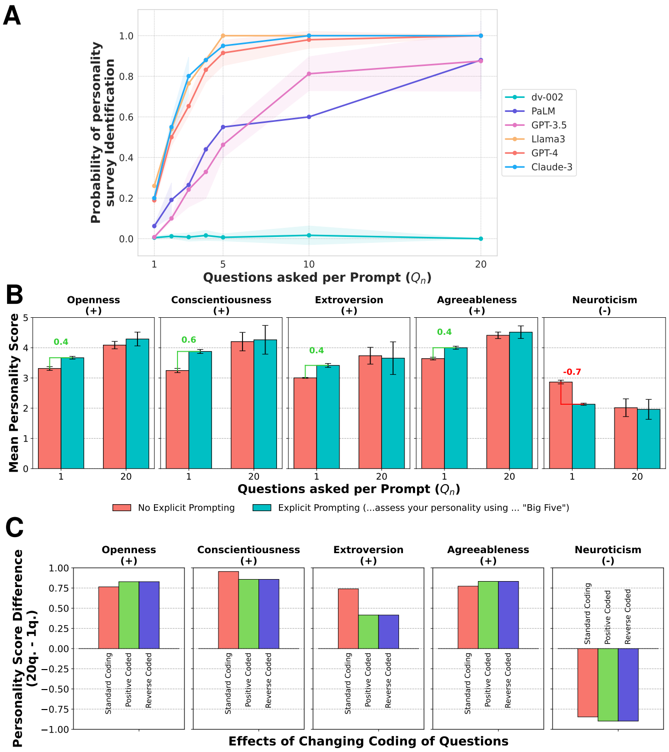


Figure 2: (A) Comparing LLMs' ability to identify the source of questions as a personality survey as a function of number of questions. (B) Big Five scores for GPT-4 with and without explicitly prompting that the LLM is completing a Big Five personality survey. The no-explicit-prompting condition is 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 (IPIP) 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. Reverse-coding items weakened but did not eliminate bias, highlighting the need for further research into the development of mitigation strategies. Our findings demonstrate potential subtleties with using psychometric tests to evaluate LLM behavior and raise concerns around using LLMs to simulate human participants.

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 [6] 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.

There are a few tactics that may mitigate response biases in simulated human data derived from LLMs. Reverse-coding items may be the most straightforward remedy; in our experiments, it reduced the bias by roughly half. Adapting surveys to be less explicitly evaluative (or using existing less-evaluative scales such as 5) may further reduce such biases. The most effortful but also most effective safeguard would be to follow classic psychometric advice [14] to replicate results across multiple measures (e.g., self-report and behavioral data), paradigms (e.g., different experiments or study designs), and data sources (including the replication of effects in human samples).

There is no doubt that LLMs are skilled at imitating humans in manifold tasks. Whether they can also generate distributions of data that accurately reflect human psychology within and across cultures remains in doubt [15]. We show evidence of response biases to a common personality assessment and provide recommendations for remediating these biases. Simulated human data from LLMs have the potential to profoundly expand our ability to carry out psychological experiments—but that will only be possible once systematic influences impacting LLM responses are well understood.

Methods

LLM Configuration We used a suite of widely available models: OpenAI's text-davinci-002, GPT-3.5, and GPT-4 (the 0613 fixed versions), Anthropic's Claude 3 Opus (claude-3-opus-20240229), Claude 3 Haiku (claude-3-haiku-20240229), Google's PaLM 2 (chat-bison-001), and Meta's Llama 3 70B Instruct, Llama 2 70B Chat. All experiments were carried out August 1-December 1, 2023. Default settings were maintained for all models, except for the temperature parameter, which was varied systematically.

Variation of Temperature Parameter We assessed the impact of the LLMs' stochastic output on response biases by varying the temperature parameter at 0.0, 0.4, 0.8, and 1.2. This range allowed us to demonstrate the robustness of the bias across different levels of response randomness.

Paraphrasing and Randomization Paraphrased variants of the survey were manually created to maintain semantic integrity while altering phrasing. We employed three randomization approaches – complete randomization of all items, randomization of questions within each factor, and no randomization (i.e., administer the items in their original sequence from the International Personality Item Pool).

Questionnaire Administration to LLMs The Big Five personality assessment was Goldberg's (1999) IPIP representation of Costa and McCrae's NEO-PI-R Domains [16], comprising 100 items scored on a 5-point Likert scale (Supplementary Information section SI.3.). LLMs received instructions akin to those provided to human participants. The only adaptation to the instructions for LLMs was to explicitly ask it to respond with exactly Q_n numbers corresponding to the batch size. The Supplementary Information (section SI.1. and SI.2.) detail the exact prompts verbatim. We dynamically replaced {{Qn}} and {{Survey_Items}} with the batch size and the corresponding survey items respectively.

Data and Code 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] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models," *arXiv:2206.07682*, 2022.

[2] G. V. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," in *International Conference on Machine Learning*, pp. 337–371, PMLR, 2023.

[3] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[4] D. Dillion, N. Tandon, Y. Gu, and K. Gray, "Can ai language models replace human participants?," *Trends in Cognitive Sciences*, 2023.

[5] J. K. Wood, J. Anglim, and S. Horwood, "A less evaluative measure of big five personality: Comparison of structure and criterion validity," *European Journal of Personality*, vol. 36, no. 5, pp. 809–824, 2022.

[6] M. Safdari, G. Serapio-García, C. Crepy, S. Fitz, P. Romero, L. Sun, M. Abdulhai, A. Faust, and M. Matarić, "Personality traits in large language models," *arXiv preprint arXiv:2307.00184*, 2023.

[7] Q. Mei, Y. Xie, W. Yuan, and M. O. Jackson, "A turing test of whether ai chatbots are behaviorally similar to humans," *Proceedings of the National Academy of Sciences*, vol. 121, no. 9, p. e2313925121, 2024.

[8] N. Schwarz, "Self-reports: How the questions shape the answers.," *American psychologist*, vol. 54, no. 2, p. 93, 1999.

[9] M. Binz and E. Schulz, "Using cognitive psychology to understand gpt-3," *Proceedings of the National Academy of Sciences*, vol. 120, no. 6, p. e2218523120, 2023.

[10] V. Dentella, F. Günther, and E. Leivada, "Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias," *Proceedings of the National Academy of Sciences*, vol. 120, no. 51, p. e2309583120, 2023.

[11] T. Holtgraves, "Social desirability and self-reports: Testing models of socially desirable responding," *Personality and Social Psychology Bulletin*, vol. 30, no. 2, pp. 161–172, 2004.

[12] B. T. Hughes, C. K. Costello, J. Pearman, P. Razavi, C. Bedford-Petersen, R. M. Ludwig, and S. Srivastava, "The big five across socioeconomic status: Measurement invariance, relationships, and age trends," *Collabra: Psychology*, vol. 7, no. 1, p. 24431, 2021.

[13] D. M. Dueber, M. D. Toland, J. E. Lingat, A. M. Love, C. Qiu, R. Wu, and A. V. Brown, "To reverse item orientation or not to reverse item orientation, that is the question," *Assessment*, vol. 29, no. 7, pp. 1422–1440, 2022.

[14] D. T. Campbell and D. W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix.," *Psychological bulletin*, vol. 56, no. 2, p. 81, 1959.

[15] S. Havaldar, B. Singhal, S. Rai, L. Liu, S. C. Guntuku, and L. Ungar, "Multilingual language models are not multicultural: A case study in emotion," in *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (J. Barnes, O. De Clercq, and R. Klinger, eds.), (Toronto, Canada), pp. 202–214, Association for Computational Linguistics, July 2023.

[16] L. 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*, vol. 7, no. 1, pp. 7–28, 1999.