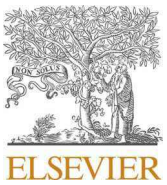
Personality and Individual Differences 217 (2024) 112434



Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Short Communication

Can a computer outfake a human?

Jane Phillips *, Chet Robie

Wilfrid Laurier University, Lazaridis School of Business & Economics, 75 University Avenue West, Waterloo, ON N2L 3C7, Canada

ARTICLE INFO

Keywords:
Personality
Single stimulus
Forced choice
Generative AI
Large language models

ABSTRACT

Faking on personality tests continues to be a challenge in hiring practices, and with the increased accessibility to free, generative AI large language models (LLM), the difference between human and algorithmic responses is difficult to distinguish. Four LLMs–GPT-3.5, Jasper, Google Bard, and GPT-4 were prompted to provide ideal responses to personality measures, specific to a provided job description. Responses collected from the LLM's were compared to a previously collected student population sample who were also directed to respond in a ideal fashion to the same job description. Overall, score comparisons indicate the superior performance of GPT-4 on both the single stimulus and forced-choice personality assessments and reinforce the need to consider more advanced options in preventing faking on personality assessments. Additionally, results from this study indicate the need for future research, especially as generative AI improves and becomes more accessible to a range of candidates.

1. Introduction

Large language models (LLM's) like ChatGPT and other artificial intelligence (AI) models are gaining an increased amount of interest (Budhwar et al., 2023) and speculation (Bodroza et al., 2023) in applied psychology and HRM. ChatGPT is significantly different than previous AI algorithms because it generates recommendations based solely off prompts from users, also known as generative AI (Budhwar et al., 2023). As LLM's continue to pull from large data sources, previous claims of their human-like intelligence are not far off (Kelan, 2023). As accessibility to these resources increase, more people are using LLM's for assistance with writing, predictions, and other tasks in a variety of fields. However, the full consequences are largely unknown and uncertain (Budhwar et al., 2023). And they continue to evolve from 2018 AI with GPT to now LLMs, with ChatGPT-4 in 2023 (Budhwar et al., 2023). Despite growing speculation for the full impact of using LLM's, they are currently used in hiring by large companies to assess high volume candidate text responses for desired verbiage and tonality (Kelan, 2023).

In efforts to understand the candidates in selection, personality measures continue to be a popular way to understand the applicant and their fit for a position (Christiansen et al., 2005). To address an ongoing need for improving self-reported personality measures (Fuechtenhans & Brown, 2022), forced-choice (FC) items have been established as better indicators of personality measures than single stimulus (SS) items (Christiansen et al., 2005) for high-stakes selection. Additionally, there is an increasing concern for distortion in self-directed responses for personality measures (Christiansen et al., 2005), which increases parallel to an emphasis on success (Fuechtenhans & Brown, 2022). When 30 to 50 % of survey participants indicate that they have misrepresented themselves during the processes of hiring and selection (Fuechtenhans & Brown, 2022), it is no surprise that 70 % of practitioners indicate that they prefer to use personality measures with some sort of faking prevention (Christiansen et al., 2005). Even though SS is currently the most common format for responding, intentionally altered responses occur more often in SS, making FC a greater opportunity for personality assessment (Christiansen et al., 2005).

Some studies have examined the personality of LLM's like ChatGPT-3.5 using a variety of instruments like Big Five Inventory-2 and HEXACO-100, finding temporal reliability and high social desirability (Bodroza et al., 2023). But, with a wide variety of stakeholders collectively building and improving generative AI, some influencing groups are becoming more dominant (Kelan, 2023), and further research is required. LLM's can be used to generate literature reviews, and although high in plagiarism, can clearly compile a comprehensive detail (Aydın & Karaarslan, 2022). ChatGPT can generally perform in recognizing personality in RNN, RoBERTa and HPMN models through text and is effective in personality prediction (Ji et al., 2023). However, more development is required as LLM's have also exhibited unfair treatment of certain groups of marginalized demographics (Ji et al., 2023). In AI supported hiring, there have been challenges in removing the

* Corresponding author.
E-mail addresses: phil2420@mylaurier.ca (J. Phillips), crobie@wlu.ca (C. Robie).

https://doi.org/10.1016/j.paid.2023.112434
Received 30 July 2023; Received in revised form 27 September 2023; Accepted 2 October 2023
Available online 6 October 2023

perpetuation of bias (algorithmic bias). In 2023, Kelan introduced the concept of bias generated from notions of fit or inflexible fixed categories for job requirements in AI (Kelan, 2023). And the superior accuracy of simple models (like Borda counts) over machine learning (ML) has been illustrated when making predications for candidate selection (Harman & Scheuerman, 2023).

Previous studies have looked at how AI is used in selection, but understanding is limited for how LLM's are being used by those faking when applying for employment. With free, publicly accessible LLM's, a candidate could easily cut and paste selection questions to best fake their fit within an organization. The current study's aim is to measure the faking abilities of ChatGPT-3.5, ChatGPT-4, Jasper, and Google Bard against a previously collected student sample, to compare the performance of human participants and generative AI.

2. Material and methods

Archival data from a recently collected data set was used for the university student sample. Participants were recruited from undergraduate university subject pools in four universities (two from Canada and two from the US) ($N = 869$). The original design required students to answer honestly at Time 1 and respond as if they were applying for a sales position (and were given a sales job description for context) at Time 2. For the purposes of the present study, only the Time 2 data were used. Extraversion and conscientiousness were the traits deemed most essential for the job given previous meta-analytic work (cf. Barrick & Mount, 1991).

Participants completed two personality assessments that used the same pool of adjectives but differed in design. We included both single stimulus and forced-choice assessments because research has found forced-choice assessments to be less fakable (Cao & Drasgow, 2019). Each personality assessment measured the FFM-based traits of extraversion, conscientiousness, agreeableness, and openness. One personality assessment was single stimulus with each of the four scales containing 20 adjectives with participants reporting their endorsement on a seven-point unipolar format ("very untrue of me" to "very true of me"). The other personality assessment was a 40 pair forced-choice assessment which required participants to choose which of the two desirability-matched adjectives described them best. We only report scores from extraversion and conscientiousness given that these are the target traits for the sales position. Details on these assessments and the faking induction can be found in Christiansen et al. (2005).

We collected data using these same personality assessments in June 2023 from four LLM's (ChatGPT-3.5, ChatGPT-4, Jasper, and Google Bard). Initially, the LLM's were asked: Can you help me choose the most appropriate option based on a job description? Then, LLM's were provided with the job description and questions from the personality measure which were individually pasted into the text field of each LLM. When required, additional prompts were used to ensure that they LLM provided appropriate responses (e.g., "you must choose one"). Testing was done on multiple occasions, but in one session for each LLM. ChatGPT-4 was tested 25 questions at a time, due to OpenAI restrictions.

3. Results

Data was analyzed using two-sided one sample *t*-tests with the LLMs acting as population parameters. Associated Cohen's *d*s (with 95 % confidence intervals) were also computed. Finally, the percentile that a given LLM achieved referencing the student sample distribution was computed. Each LLM was compared independently to the student sample. All scores were converted to a 100-point scale using simple linear interpolations to aid in comparability.

Means, standard deviations, *t*-tests, *d* scores (and 95 % confidence intervals), and percentiles can be found in Table 1. Results are graphically illustrated in Fig. 1. As can be seen from Fig. 1, most of the LLMs performed at or above the median of the student population in faking the single stimulus measure (with the exception of Google Bard for Extraversion which fell at the 33rd percentile of the student population). However, ChapGPT-4 clearly performed the best scoring at the 99th and 100th percentile of the student sample for extraversion and conscientiousness, respectively. Most of the LLMs had greater trouble in faking the forced-choice assessment with most falling at or below the median of the student population. However, again, ChatGPT-4 performed the best with percentiles of 85.2 and 98.4 referencing the student population for extraversion and conscientiousness, respectively.

4. Discussion

The aim of this study was to compare SS and FC personality scores, prompted for faking, generated from four different LLM's (GPT-3.5, Jasper, Google Bard, and GPT-4) against a previously collected student population. The purpose of this comparison was to examine whether the generative AI responses, when prompted to fake, would be more successful than the student population. Results indicate that SS Likert-type questions were easier to fake, whereas FC were more difficult which is consistent with research on human populations (Cao & Drasgow, 2019). Additionally, LLM's had a harder time faking conscientiousness. This may be attributed to an emphasis extraversion, as a more stereotypical key trait for the job description. Connecting conscientiousness to the provided job description would be more subtle.

Although the LLM's had varied results, GPT-4 outperformed most of the student population, faking on average better than 99.6 % of the

Table 1
Comparison of faking ability of four LLMs to a sample of university students.

LLM	Extraversion								Conscientiousness							
						95 % CI								95 % CI		
	M_1	M_2	SD_2	t	d	d_{ll}	d_{ul}	%tile	M_1	M_2	SD_2	t	d	d_{ll}	d_{ul}	%tile
Single stimulus																
ChatGPT 3.5	77.5	65.1	15.8	**23.2**	0.79	0.65	0.93	76.1	85.8	77.4	13.1	**19.0**	0.65	0.51	0.78	69.9
Jasper	67.5	65.1	15.8	**4.5**	0.15	0.02	0.29	54.2	76.7	77.4	13.1	−1.6	−0.06	−0.19	0.08	44.2
Google Bard	56.7	65.1	15.8	**−15.7**	−0.53	−0.67	−0.40	32.7	83.3	77.4	13.1	**13.4**	0.45	0.32	0.59	66.2
ChatGPT-4	95.8	65.1	15.8	**57.4**	1.95	1.79	2.11	99.2	100.0	77.4	13.1	**51.0**	1.73	1.57	1.88	100.0
Forced-choice																
ChatGPT 3.5	31.6	45.1	22.0	**−18.2**	−0.62	−0.75	−0.48	32.6	52.4	52.6	19.5	−0.3	−0.01	−0.14	0.12	45.2
Jasper	31.6	45.1	22.0	**−18.2**	−0.62	−0.75	−0.48	32.6	52.4	52.6	19.5	−0.3	−0.01	−0.14	0.12	45.2
Google Bard	42.1	45.1	22.0	**−4.1**	−0.14	−0.27	0.00	50.7	61.9	52.6	19.5	**14.1**	0.48	0.35	0.62	64.7
ChatGPT-4	73.7	45.1	22.0	**38.2**	1.30	1.15	1.44	85.2	90.5	52.6	19.5	**57.4**	1.95	1.79	2.11	98.4

Note. Student $N = 869$. LLM = language learning model. All scale scores have been converted to a 0 to 100 metric using simple linear transformations. M_1 = mean LLM scale score. M_2 = mean student scale score. SD_2 = standard deviation of student scale score. d = Cohen's *d*. 95 % CI = 95 % confidence interval. LL = lower limit. UL = upper limit. %tile = percentile at which LLM scale score falls in the student scale score distribution. Bold t values denote $p < .001$ for the one-sample *t*-tests.

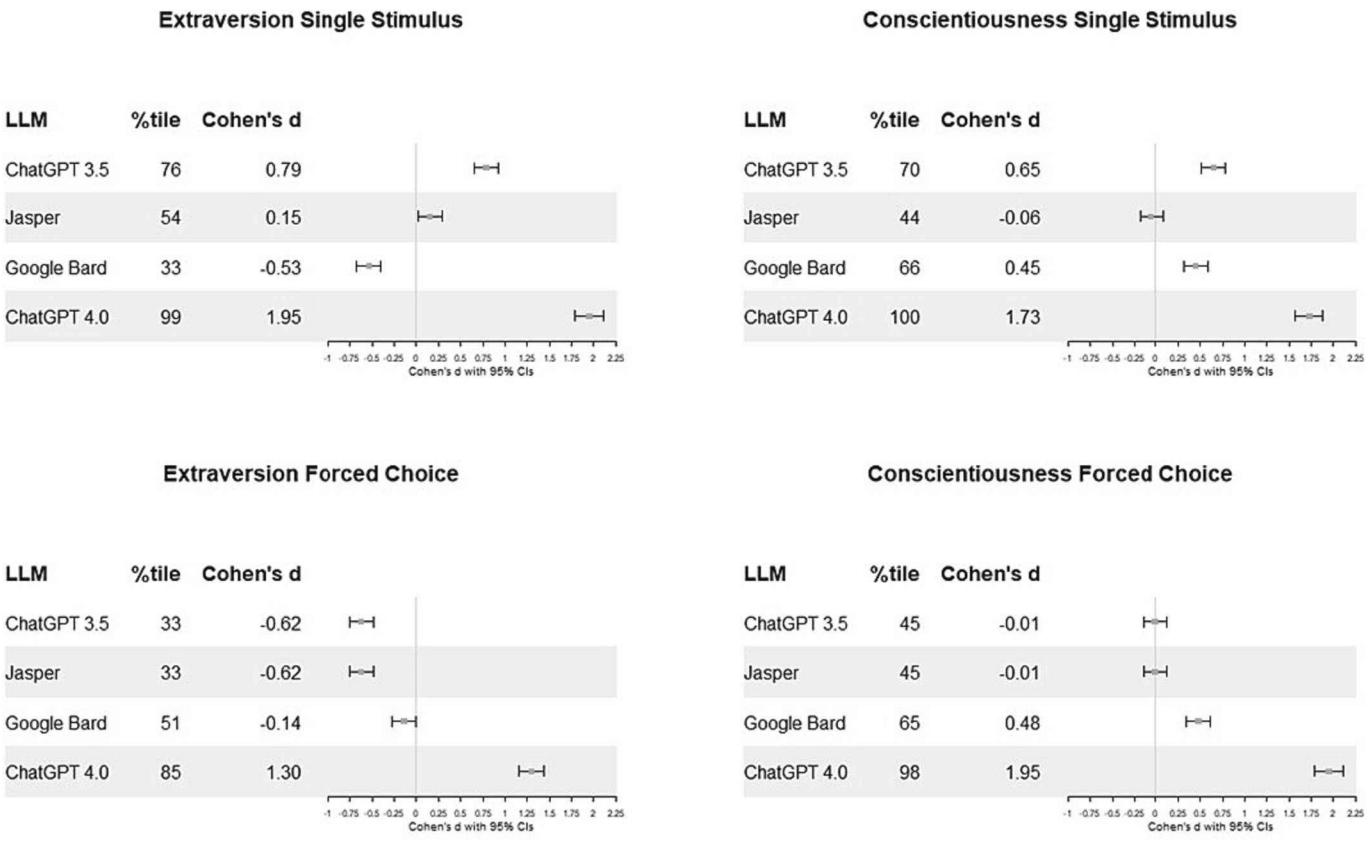


Fig. 1. Graphical illustrations comparing faking ability of four LLMs to a sample of University Students.

student population on Likert-type measures, and 91.78 % better than the student population on the forced-choice measures. Single stimulus assessments are relatively easy to fake; whereas practitioners often opt for forced-choice assessments to combat faking. The primary practical implication of these results is that generative AI may soon make prevention of faking on noncognitive assessments in personnel selection much more difficult. Further testing is needed to see how the LLM's responses would change longitudinally, as generative AI is improved with additional resources (i.e., phrase-based learning, construct variety). Future research should also examine whether LLMs have more difficulty faking phrase-based forced-choice assessments with more than two options per block or whether speeding assessments (e.g., rapid response measurement) may defeat generative AI assistance (Meade et al., 2020). Additional research should also further test for the variability of answers from the LLM's when issued in the same and different chats.

This study was limited by LLM availability. Only publicly accessible LLM's were used in this study. Paid or beta LLM's may yield significantly better results. This limitation was put in place to test publicly accessible LLM's for the general population. Additionally, all data was collected in June 2023. Since technology is constantly changing and improving, this may influence the timeliness of this data.

5. Conclusion

The results of this study indicate the potential effectiveness of LLM's in assistance with faking personality measures for SS and FC items. Although GPT-3.5, Jasper, and Google Bard had varied results, GPT-4 significantly outperformed the other LLM's and previously collected student population. Future research is needed to continue developing an understanding of the impact of generative AI and LLM's in candidate faking opportunities on personality measures.

CRediT authorship contribution statement

J. Jane Phillips: Conceptualization, Methodology, Analysis, Writing.

Chet Robie: Methodology, Analysis, Writing.

Data availability

Data will be made available on request.

References

Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. In Ö. Aydın (Ed.), *2. Emerging computer technologies* (pp. 22–31). İzmir Akademi Dernegi.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Bodroza, B., Dinic, B., & Bojic, L. (2023). Personality testing of GPT-3: Limited temporal reliability but highlighted social desirability of GPT-3's personality instruments results. *Computer Science, Artificial Intelligence*. https://doi.org/10.48550/arXiv.2306.04308

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G., Beltran, J., … …Varma, A.. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal, 33*(3), 606–659. https://doi.org/10.1111/1748-8583.12524

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

Fuechtenhans, M., & Brown, A. (2022). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment, 31*(1), 105–119. https://doi.org/10.1111/ijsa.12409

Harman, J. L., & Scheuerman, J. (2023). Simple rules outperform machine learning for personnel selection: Insights from the 3rd annual SIOP machine learning

competition. *Discover Artificial Intelligence, 3*(2). https://doi.org/10.1007/s44163-022-00044-2

Ji, Y., Wu, W., Zheng, H., Hu, Y., Chen, X., & He, L. (2023). Is ChatGPT a good personality recognizer? A preliminary study [Unpublished manuscript]. https://arxiv.org/pdf/2307.03952.pdf.

Kelan, E. K. (2023). Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring. *Human Resource Management Journal.*. https://doi.org/10.1111/1748-8583.12511

Meade, A. W., Pappalardo, G., Braddy, P. W., & Fleenor, J. W. (2020). Rapid response measurement: Development of a faking-resistant assessment method for personality. *Organizational Research Methods, 23*(1), 181–207. https://doi.org/10.1177/1094428118795295