



Delft University of Technology

The use of ChatGPT for personality research Administering questionnaires using generated personas

de Winter, Joost C.F.; Driessen, T.; Dodou, Dimitra

DOI

[10.1016/j.paid.2024.112729](https://doi.org/10.1016/j.paid.2024.112729)

Publication date

2024

Document Version

Final published version

Published in

Personality and Individual Differences

Citation (APA)

de Winter, J. C. F., Driessen, T., & Dodou, D. (2024). The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, 228, Article 112729. <https://doi.org/10.1016/j.paid.2024.112729>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



The use of ChatGPT for personality research: Administering questionnaires using generated personas

Joost C.F. de Winter^{*}, Tom Driessen, Dimitra Dodou

Faculty of Mechanical Engineering, Delft University of Technology, Netherlands

ARTICLE INFO

Keywords:

Large language models
Big Five
Personas
Personality research

ABSTRACT

Personality research has traditionally relied on questionnaires, which bring with them inherent limitations, such as response style bias. With the emergence of large language models such as ChatGPT, the question arises as to what extent these models can be used in personality research. In this study, ChatGPT (GPT-4) generated 2000 text-based personas. Next, for each persona, ChatGPT completed a short form of the Big Five Inventory (BFI-10), the Brief Sensation Seeking Scale (BSSS), and a Short Dark Triad (SD3). The mean scores on the BFI-10 items were found to correlate strongly with means from previously published research, and principal component analysis revealed a clear five-component structure. Certain relationships between traits, such as a negative correlation between the age of the persona and the BSSS score, were clearly interpretable, while some other correlations diverged from the literature. An additional analysis using four new sets of 2000 personas each, including a set of 'realistic' personas and a set of cinematic personas, showed that the correlation matrix among personality constructs was affected by the persona set. It is concluded that evaluating questionnaires and research hypotheses prior to engaging with real individuals holds promise.

1. Introduction

Research into personality primarily relies on questionnaires (Bau-meister et al., 2007). Personality traits obtained in this manner have demonstrated criterion validity, as demonstrated by correlations with job performance (Barrick & Mount, 1991), health-related behaviors (Bogg & Roberts, 2004), consumer behavior (Matz et al., 2016), and academic performance (Poropat, 2009), among others.

At the same time, questionnaires are known to exhibit various limitations. First, there are ethical factors to account for. Administering questionnaires demands time commitments from respondents, and some questions may be regarded as sensitive and therefore yield non-responses. Second, human respondents have imperfect reliability. Repeating the same questionnaire to improve statistical reliability is often unviable due to learning, adaptation, boredom, or fatigue. Third, questionnaire research is known to be plagued by socially desirable answers and other biases (e.g., Tourangeau & Yan, 2007). Research has shown disparities between self-assessments of personality and evaluations of the same individuals' personalities by friends or acquaintances (e.g., Moritz & Roberts, 2018). Such disparities might arise from the limited accuracy with which individuals can appraise themselves or

others. However, such differences might also be attributed to response styles (e.g., Biderman et al., 2011). The latter topic has spurred considerable debate, for example, regarding whether the Big One is a valid construct or merely a by-product of common method variance (Arias et al., 2018; Davies et al., 2015; Just, 2011).

Recently, there has been a surge of interest in large language models (LLMs), with ChatGPT being the most well-known example. While ChatGPT outputs can exhibit biases and generate inaccurate information (Bubeck et al., 2023), ChatGPT is known to excel in deciphering linguistic structures. In certain benchmark tests, such as the Graduate Record Examinations (GRE) Verbal Section as well as reading comprehension tasks, ChatGPT-4 shows particularly strong performance (De Winter, 2023; OpenAI, 2023).

According to the lexical hypothesis, relevant adjectives that describe human traits are historically embedded in everyday language and are present in modern dictionaries (Cattell, 1943; John et al., 1988). The Big Five model was developed from these adjectives through statistical reduction (Goldberg, 1990; Saucier & Goldberg, 1996). The close linkage between personality and language raises the question of whether ChatGPT has potential applicability in personality research.

Several studies have examined the potential of ChatGPT for

^{*} Corresponding author.

E-mail address: j.c.f.dewinter@tudelft.nl (J.C.F. de Winter).

personality-related research. For example, ChatGPT has been found to be proficient in estimating personality from textual descriptions, text messages, or digital footprints (Cao & Kosinski, 2024; Ji et al., 2023; Peters & Matz, 2023; Rao et al., 2023). Phillips and Robie (2024) showed that LLMs could produce tailored responses to a personality questionnaire in order to fit a given job description (i.e., test faking). Furthermore, ChatGPT is widely used to mimic certain professions or skills. Specifically, role-playing prompting (e.g., ‘you are a useful assistant’, ‘you are a world-class Python programmer’) is a commonly used technique to have ChatGPT adopt a certain persona, and thus obtain more useful answers (e.g., Kong et al., 2024). Another application area for LLMs is the simulation of agents that possess diverse types of knowledge and personalities. By letting such agents interact, plausible behaviors of those agents have been shown to emerge in a microworld (Park et al., 2023). Wang et al. (2024) suggested that team performance of such LLM-based agents improves when team members are configured with different personal characteristics and expertises, similar to how human team members can cognitively complement each other.

Other studies have studied the use of synthetic humans (personas) to fill out questionnaires using LLMs (Argyle et al., 2023; Bisbee et al., 2024; Olea et al., 2024; Serapio-García et al., 2023), with overall promising results. In particular, Serapio-García et al. (2023) used text-based personas and let diverse LLMs (PaLM variants) complete various personality scales: the International Personality Item Pool-Neuroticism, Extraversion, Openness (IPIP-NEO), Big-Five Inventory (BFI), Positive and Negative Affect Schedule Scales (PANAS), Buss-Perry Aggression Questionnaire (BPAQ), Revised Portrait Values Questionnaire (PVQ-RR), and Short Scale of Creative Self (SCSS). The authors reported convergent validity (high correlations for the same constructs measured using the IPIP-NEO and the BFI), discriminant validity (differences in correlations for different constructs measured using the IPIP-NEO versus the BFI), and criterion validity (expected correlations between the IPIP-NEO and the other tests). However, more attention should be paid to the validity of results obtained through synthetic personas with respect to the published literature.

The research conducted to date on personas that complete questionnaires using an LLM is insightful and offers various possible applications, such as enabling more realistic interactions with chatbots or agents in virtual worlds (e.g., Klinkert et al., 2024). However, more research is needed in the area of personas and personality from a psychometric perspective. The current study seeks to determine whether the underlying dimensions of the BFI-10 can be derived through personas, the extent to which the dimensions of the Big Five personality dimensions correlate with other personality dimensions, and whether these associations correspond with the published literature. We also examined the extent to which the type of persona (default, created to be more realistic, or more like a film character) influenced the nomological network of correlations.

Building upon previous studies, such as those in which ChatGPT was used to rate the quality of scientific abstracts (De Winter, 2024), we propose a pipeline that involves the generation of personas. For these personas, we had ChatGPT complete a Big Five personality test as well as four additional personality scales, and we compared correlations with existing literature obtained using human respondents. Finally, we discuss potential applications and limitations of using LLMs to simulate human personas.

2. Methods and results

First, we generated 2000 personas using the ChatGPT API (model gpt-4-0613; 28 and 29 January 2024). The temperature parameter in the ChatGPT API adjusts the probability distribution of the next token to be produced, with a minimum setting of 0 for nearly deterministic outputs and a maximum allowed value of 2 for highly random outputs. We used a temperature setting of 1, meaning that some randomness was introduced. We used the following prompt:

Prompt:

Please give me 20 personas.

Also add their age and gender. The persona should be described by means of three brief sentences separated by semicolons.

Report each persona on a single line, numbered 0001 to 0020. Separate age, gender, profession/activity/job versus description by means of a dash. Only personas; nothing else.

We applied this in 100 batches, with the message history of each batch containing the original prompt and the last 20 personas. A new batch was generated using the following prompt:

Prompt:

Please generate 20 more personas. Continue counting, from 0021 to 0040

The above prompt yielded personas with the following types of descriptions:

ChatGPT-4 output:

0001. 24, Female, Graphic Designer - Always has her sketchbook at hand; creative and innovative; obsessed with the latest design trends.

0002. 32, Male, Software Engineer - Quiet person who loves puzzles; spends free time coding and playing video games; always eager to learn something new.

0003. 39, Female, Yoga instructor - Dedicated to promoting health and wellbeing through physical activity; spends most of her day in the yoga studio; enjoys hiking and the outdoors.

...

2000. 31, Female, Content Creator - Manages her own lifestyle blog; collaborations with lifestyle and beauty brands; loves yoga and landscape photography.

Next, we had each persona complete a short version of the Big Five Inventory (BFI-10; Rammstedt & John, 2007). To do this, we programmatically went through all individual personas, and presented the BFI-10, using the following prompt:

Prompt:

Imagine the following person (age, gender, description): *24, Female, Graphic Designer - Always has her sketchbook at hand; creative and innovative; obsessed with the latest design trends.*

Rate this person / complete the questionnaire for this person, on a scale of 1 (disagree strongly) to 5 (agree strongly).

Report in the following format on a single line, e.g., 5 1 4 ...

Report only digits, nothing else.

1. gets nervous easily.

2. tends to find fault with others.

3. is outgoing, sociable.

4. is generally trusting.

5. tends to be lazy.

6. is relaxed, handles stress well.

7. has few artistic interests.

8. does a thorough job.

9. is reserved.

10. has an active imagination.

In the above prompt, the persona is highlighted in Italics; the prompt submitted to the API was in plain text only. In each prompt, the Italics phrase was replaced with another persona. Each prompt was submitted independently, without using a chatting functionality. The gpt-4-turbo-2024-04-09 model was used (date range of prompting: 30 April – 4 May

2024). The temperature parameter was set to 0, which means that we opted for the most deterministic setting of the ChatGPT API. Each persona was prompted 10 times. The 10 items in the prompt were sorted in a random order for each individual prompt, based on prior recommendations (Rao et al., 2023; Tabone & De Winter, 2023; Tang et al., 2024; Wang et al., 2023). This approach, known as ‘self-consistency’, was used to improve the statistical reliability of the output and to counteract potential item-order effects, a phenomenon that might have parallels with the anchoring effect in humans (Furnham & Boo, 2011).

In total, 20,000 prompts (2000 personas $\times$ 10 repetitions) were executed. Numerical outputs were unavailable in 19 out of 20,000 cases. This occurred because ChatGPT produced an output that was not in the desired format, for example by producing more digits than the expected 10. After sorting the items back into a fixed sequence, a mean score was calculated across the (usually) 10 repetitions. This resulted in a matrix of 2000 rows (corresponding to the 2000 personas) and 10 columns (corresponding to the BFI-10 items), containing numbers between 1 and 5.

This above procedure was separately applied to four other questionnaires using the same prompt layout: (1) an 8-item Brief Sensation Seeking Scale (BSSS; Hoyle et al., 2002), (2) a 9-item Machiavellianism scale, (3) a 9-item narcissism scale, and (4) a 9-item psychopathy scale (Jones & Paulhus, 2014). These four additional scales were administered because they are widely used and are conceptually distinct from the BFI, making them suitable for assessing how their scores correlate with the BFI scores.

After administering all five questionnaires, we performed a principal component analysis (PCA) on the BFI-10 results, followed by oblique Promax rotation of the loadings (Hendrickson & White, 1964). Finally, the components of the rotated component loading matrix were ordered and sign-reversed if needed, in order to yield a maximal congruence coefficient (MacCallum et al., 1999; Tucker, 1955) with respect to a target matrix consisting of 0s and 1s.

The means, standard deviations, and loadings of the BFI-10 are shown in Table 1. The component loadings have the expected sign, with positive and negative loadings for the two items of the same personality dimension. The loading matrix can be seen to have a highly simple structure, where simple structure is defined as a loading matrix where each item has a strong loading on one component and weak loadings on the other components (e.g., Revelle & Rocklin, 1979; Thurstone, 1935; Tucker, 1955).

The listed means show lower values for negative characteristics (e.g., ‘tends to be lazy’), consistent with the tendency of people to assess themselves with positive regard. The Pearson product-moment correlation of the item means with the corresponding item means from a previously published BFI-10 study (Lovik et al., 2017) was strong, at $r = 0.93$ ($p < 0.001$). This association is illustrated in Fig. 1. It can also be noted that the averages deviated substantially, with ChatGPT being more inclined than human raters to report extreme values (i.e., close to 1

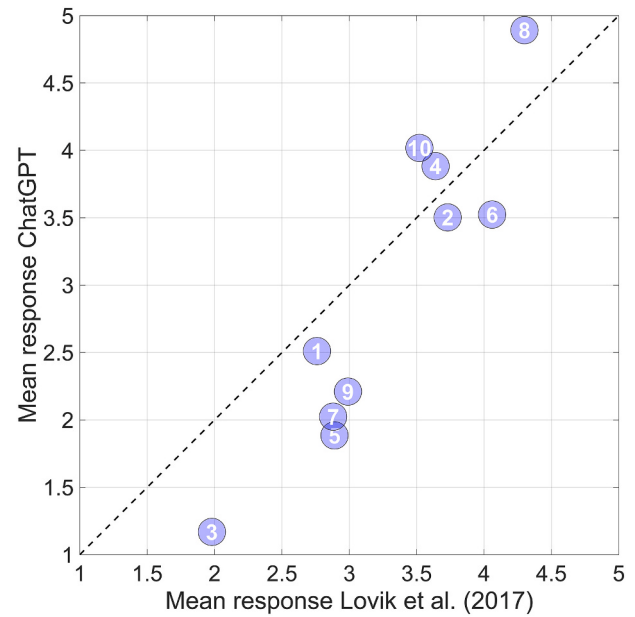


Fig. 1. Mean response to the BFI-10 by ChatGPT and as reported by Lovik et al. (2017). The number in each marker represents the item number (see Table 1).

or 5).

Table 2 shows the means, standard deviations, and Pearson product-moment correlation coefficients of the scale scores, together with the age and gender of the personas, while the supplementary material (Table S1) provides an overview of corresponding correlation coefficients obtained from published literature, including meta-analyses (Aluja et al., 2022; Cross et al., 2013; Crysel et al., 2013; Dahlen & White, 2006; Denissen et al., 2008; Eachus, 2004; Jones & Paulhus, 2014; Muris et al., 2017; Schmitt et al., 2008; Van der Linden et al., 2010).

The sign of various correlations shown in Table 2 was found to correspond with the literature. For example:

- Age was positively correlated with conscientiousness ($r = 0.27$ vs. 0.24) and negatively with sensation seeking ($r = -0.48$ vs. -0.34), neuroticism ($r = -0.30$ vs. -0.15), and psychopathy ($r = -0.27$ vs. -0.19) (Aluja et al., 2022; Denissen et al., 2008).
- The positive associations between sensation seeking, Machiavellianism, narcissism, and psychopathy are consistent with the literature (Crysel et al., 2013; Jones & Paulhus, 2014).
- Also consistent with the literature is the positive association between extraversion and sensation seeking ($r = 0.59$ vs. 0.31) (Dahlen &

Table 1

Means, standard deviations (SD), and component loadings for the Big Five Inventory (BFI-10) ($n = 2000$).

Item	Trait	Mean	SD	E	A	C	N	O
1. is reserved	Extraversion (R)	2.51	0.55	-0.97	-0.03	-0.03	-0.02	-0.05
2. is generally trusting	Agreeableness	3.50	0.34	0.10	0.85	0.06	0.01	-0.08
3. tends to be lazy	Conscientiousness (R)	1.17	0.25	-0.11	0.03	-0.93	0.07	0.02
4. is relaxed, handles stress well	Neuroticism (R)	3.88	0.38	-0.03	0.21	-0.12	-0.84	0.00
5. has few artistic interests	Openness (R)	1.89	0.63	-0.12	-0.04	-0.03	0.03	-0.96
6. is outgoing, sociable	Extraversion	3.53	0.78	0.97	-0.02	-0.04	0.01	-0.03
7. tends to find fault with others	Agreeableness (R)	2.02	0.24	0.09	-0.89	0.02	-0.02	-0.07
8. does a thorough job	Conscientiousness	4.89	0.23	-0.15	0.06	0.88	0.05	0.02
9. gets nervous easily	Neuroticism	2.21	0.28	0.00	0.15	-0.11	0.96	0.00
10. has an active imagination	Openness	4.02	0.74	-0.10	-0.04	-0.03	0.03	0.94

Note. The values are color-coded from -1 (orange) to 0 (white) to 1 (green).

Table 2Pearson product-moment correlation matrix among age, gender, and personality traits ($n = 2000$).

	Mean	SD	1	2	3	4	5	6	7	8	9	10
1 Age	37.45	10.12										
2 Gender (1: F, 2: M)	1.50	0.50	0.04									
3 Extraversion	7.02	1.30	-0.28	-0.10								
4 Neuroticism	4.33	0.61	-0.30	-0.04	-0.09							
5 Agreeableness	7.48	0.51	-0.02	-0.22	0.14	-0.44						
6 Openness	8.13	1.30	-0.11	-0.08	0.05	0.29	0.03					
7 Conscientiousness	9.72	0.44	0.27	-0.21	-0.26	-0.24	0.12	-0.26				
8 Sensation seeking	19.01	4.29	-0.48	0.12	0.59	0.06	-0.09	0.24	-0.39			
9 Machiavellianism	18.64	3.35	-0.02	0.15	0.11	0.41	-0.72	0.04	-0.20	0.13		
10 Narcissism	25.03	4.51	-0.10	-0.02	0.66	0.13	-0.36	0.04	-0.17	0.41	0.53	
11 Psychopathy	11.83	2.48	-0.27	0.29	0.50	0.22	-0.40	0.15	-0.52	0.68	0.52	0.55

Note. Numbers are color-coded from -0.50 (orange) to 0 (white) to 0.50 (green). The Pearson product-moment correlation with Gender, a binary variable, is equivalent to the point-biserial correlation coefficient.

White, 2006), between extraversion and narcissism ($r = 0.66$ vs. 0.31) (Muris et al., 2017), and between sensation seeking and openness ($r = 0.24$ vs. 0.20) (Dahlen & White, 2006).

- The association between agreeableness versus Machiavellianism, narcissism, and psychopathy was negative ($r = -0.72$, -0.36 , -0.40), which is consistent with literature ($r = -0.43$, -0.21 , -0.46) (Muris et al., 2017).

On the other hand, several correlations diverge from published literature:

- The literature (see Table S1) suggests that females score higher on neuroticism compared to males; however, in Table 2, females and males have equivalent neuroticism ($r = -0.04$ vs. -0.40) (Schmitt et al., 2008).
- ChatGPT indicated that older people have lower extraversion ($r = -0.28$). However, according to the literature, the effects are near-zero ($r = 0.00$) (Denissen et al., 2008).
- The correlations among the Big Five dimensions are not consistent with the literature either. For example, ChatGPT yielded a negative correlation between extraversion and conscientiousness ($r = -0.26$), while the literature reports a positive correlation ($r = 0.21$) (Van der Linden et al., 2010). Also, while the literature indicates that neuroticism correlates negatively with extraversion, agreeableness, openness, and conscientiousness ($r = -0.26$, -0.26 , -0.12 , and -0.32 , respectively), ChatGPT yielded more divergent correlations ($r = -0.09$, -0.44 , 0.29 , and -0.24 , respectively).
- Table 2 depicts a positive correlation between extraversion and psychopathy ($r = 0.50$), while the literature suggests a near-zero correlation ($r = 0.01$) (Muris et al., 2017).

In summary, the results indicate that while some correlations are consistent with previously published literature, there are also incidences of dissimilarity.

The ChatGPT base model has been trained using a wide variety of text data collected from the internet. Consequently, the personas generated by this model may not accurately represent the average citizen. To investigate whether more realistic personas could influence the correlations among constructs, we revised our original prompting strategy. Instead of requesting 'Please give me 20 personas', we used the following modified prompt:

Prompt:

Please give me 20 REALISTIC personas with REALISTIC names.

ChatGPT-4 output:

0001. John Smith, 35, Male, Software Engineer - Active in online coding communities; enjoys cycling and reading; spends weekends tinkering with new tech.

0002. Emily Harris, 27, Female, Marketing Manager - Adores fashion and beauty trends; coffee lover; attends a yoga class twice a week.

...

And

Prompt:

Please give me 20 REALISTIC personas with REALISTIC names and REALISTIC personality descriptions.

ChatGPT-4 output:

0001. Laura Richards, 30, Female, Digital Marketer - Laura is extremely driven and loves the fast-paced nature of her industry; her colleagues admire her for her creative problem-solving skills; she has an outgoing personality, but values her alone time.

0002. Henry Davis, 54, Male, Physics Teacher - Henry is known for his patience and ability to explain complex concepts in an easy manner; he's introverted and prefers to spend his time reading or gardening; he deeply cares about his students and their success.

...

We also explored whether incorporating terminology from the Big Five personality traits into the prompt would result in more accurate correlations between constructs. To test this, we generated personas using the following revised prompt:

Prompt:

Please give me 20 REALISTIC personas with REALISTIC names and REALISTIC personality descriptions using terminology of the Big Five.

ChatGPT-4 output:

0001. Jane Wilson, 25, Female, Graduate Student - Highly conscientious; scores low in emotional stability and can often feel anxious; but is very open to new experiences and ideas.

0002. Richard Meyers, 32, Male, Software Engineer - Very extroverted and tends to bond instantly with new people; presents high agreeableness; lacks emotional stability leading to occasional unexpected emotional outbreaks.

...

Finally, we considered generating extreme personas that are unrealistic, as if they came from a movie. In this way, a greater variance might be introduced, potentially making the relationships between

constructs more apparent.

Prompt:

Please give me 20 CINEMATIC personas with CINEMATIC names and CINEMATIC personality descriptions.

ChatGPT-4 output:

0001. Indiana Jones, 39, Male, Archaeologist - Ambitious and daring; relentless in pursuit of historical relics; balances academic intellect with rugged adventurism.

0002. Ellen Ripley, 35, Female, Warrant Officer - Strong-willed and intelligent; survives against incredible odds; protective and always steps up to unanticipated challenges.

...

The component loadings of the BFI, presented in the Supplementary Material (Tables S2–S5), show that the original mean absolute salient loading of 0.92 (see Table 1) remained relatively consistent with realistic personas (0.90 and 0.89, respectively; Tables S2 & S3). By explicitly incorporating Big Five elements as part of the personas, this figure rose to 0.94 (Table S4), and cinematic personas also resulted in a high mean absolute salient loading of 0.95 (Table S5). In summary, by explicitly including Big Five keywords or by generating cinematic personas, the BFI structure became more manifest.

The correlation between constructs indicated that correlations with realistic personas remained relatively unchanged compared to Table 1 (see Tables S6 & S7), while the correlations for Big Five personas (Table S8) and cinematic personas (Table S9) showed stronger deviations from the correlations presented in Table 1. For cinematic personas, there was a noteworthy increase in standard deviations among the 2000 personas for Machiavellianism ($SD = 9.98$, compared to $SD = 3.35$ in Table 1) and Psychopathy ($SD = 10.19$ compared to $SD = 2.48$ in Table 1). On average, cinematic personas were depicted as more malevolent than the ‘realistic’ personas.

3. Discussion

Questionnaires are commonly used in personality research, but have disadvantages in the form of response style bias and the limited number of repetitions that can reasonably be performed on a human participant. Due to their text comprehension abilities, LLMs are potentially promising for administering questionnaires among synthetic participants, also referred to as personas. Building on previous research (e.g., Serapio-García et al., 2023), this study explored using ChatGPT for psychometric personality research. A pipeline was created which involved the creation of text-based personas and subsequently completing questionnaires for these personas. A total of five questionnaires were completed 10 times by each persona with the response options each time in a different random order, a prompting strategy that is based on literature (e.g., Tabone & De Winter, 2023; Tang et al., 2024).

One key result was that a very clear loading structure emerged (see Table 1), without mentioning such a structure in the prompt. Although the personas occasionally used Big Five-related terminology: for example, the word ‘extroverted’ appeared in 8 of the 2000 personas, most personas did not include such language. However, we found that including explicit Big Five terminology in a new set of 2000 personas, or using cinematic personas in another set of personas, made the simple structure even more pronounced (see Tables S8 & S9 in the Supplementary Material). Note that our results do not imply that the Big Five is the only or most dominant dimensional structure within the GPT neural network; Potentially, the structure of other personality tests would also have emerged, had these personality tests been administered.

The mean scores of the BFI-10 items were found to correspond well with literature ($r = 0.93$), although the personas displayed relatively high openness and conscientiousness (see Fig. 1). Previous research indicates that ChatGPT output is of agreeable nature (Hilliard et al., 2024;

Rutinowski et al., 2024). A possible explanation for ChatGPT's positive disposition is that it has been fine-tuned with human feedback and specifically instructed to become a more user-friendly assistant. (OpenAI, 2023). Another explanation is that ChatGPT's neural network, which has been trained on human textual data, mirrors the tendency of humans to leave a positive impression on others (i.e., impression management; Rosenberg & Egbert, 2011).

3.1. Limitations

Limitations.

Several limitations should be acknowledged. A key limitation is that although the correlations between constructs appear plausible and partly align with the literature (compare Table 2 with Table S1 in the Supplementary Material), there are still important discrepancies. The challenge is that ChatGPT, like other popular LLMs, is trained on texts obtained from the internet, and is thus susceptible to the biases present in these texts (Atari et al., 2023; Grossmann et al., 2023; for a similar discussion on big data validity, see Barocas & Selbst, 2016). Additionally, the personas we generated are inherently a product of the prompts used to create them. We showed that different sets of personas yield slightly different correlations, which is not surprising since, like surveying humans, the results inevitably depend on the sampled population. The challenge for future research lies in generating new sets of personas for which the obtained correlations closely match those from the literature on the human population that is meant to be represented.

On the other hand, it can be questioned whether representing correlations from the literature should be the intention, given that human respondents are susceptible to response biases when completing surveys. In fact, the literature indicates that the moderately strong correlations between Big Five constructs might be attributed to response styles such as social desirability bias (Biderman et al., 2011). The volatility of correlations obtained from human respondents is also evident from research where correlations between BFI constructs were found to differ depending on whether respondents evaluated themselves or were evaluated by others (DeYoung, 2006; Moritz & Roberts, 2018). Hence, it is conceivable that the results shown in Tables 1 and 2 might provide a purer (a concept still to be precisely defined) representation of the relationships between personality dimensions.

A second limitation is the common availability of the BFI-10 on the internet and its inclusion in the GPT base model (this can be easily verified by using the prompt “Give me the 10 items of the BFI-10”). Therefore, there is a possibility that ChatGPT recognized the BFI-10 items, and assigned scores to the personas by memorizing stereotypical answers, without actually judging the persona. However, this may not be a major limitation, as shown by the correlations between constructs of different questionnaires, such as between Extraversion and Sensation Seeking (see Table 2). These correlations were obtained through independent prompts, and cannot have arisen through memorization. Nevertheless, for future research, it might be interesting to repeat the current study with new questionnaires that have not yet been published on the internet.

A third limitation is that the research presented here is quite cumbersome and not yet easily applicable. With faster inference methods and lower costs, the method may become more usable. To provide an estimate: For this study, we created a total of 10,000 personas (distributed across 5 sets of 2000 each), with each persona completing 5 questionnaires 10 times. These half a million prompts cost us several days to execute using parallel prompting on a single laptop computer, with OpenAI charging us about \$1500 for executing these prompts. For faster screening of prospective questionnaires, it would be useful if this could be done more efficiently.

3.2. Suggestions for follow-up research and applications

Although our research has several limitations, it may hold promise

for follow-up research and applications. After further refinement and validation, and faster inference times of the LLMs, we see the following possible future uses of ChatGPT in future personality research:

- **Questionnaire construction:** It may be possible to assess latent structures before applying the questionnaire to real people. This includes testing of new items, reformulations of items, or removal of bad or redundant items.
- **Training, education, and hypothesis forming:** This study showed that the obtained correlation matrix is dependent on the personas. In the real world, such matters may be difficult to study without spending significant resources. For example, when comparing the traits of individuals across different countries, one is confronted with the challenge that not only is the culture different, the mean age of the respondents, as well as the mean age of the population in that country, also varies (De Winter & Dodou, 2016; Spector et al., 2015). It is conceivable that when samples are difficult to recruit, hypotheses might be preliminarily tested by relying on synthetic data, as was done in this paper. The current findings also elucidated that there is no single ‘population value’, but that the input (i.e., the type of persona) determines the output. These insights may prove valuable for statistics teaching.
- **Further exploration of response styles.** A common criticism in personality research pertains to the issue of common method variance (Biderman et al., 2011; Spector & Brannick, 2009). Through computer simulations such as ours, it should be feasible to manipulate response styles and test the effects, for example, by posing the same questions in different ways (e.g., forced-choice vs. multiple choice), with different persona perspectives (e.g., ‘imagine you are this persona’ vs. ‘fill out the questionnaire for this persona’), and in varying sequences. By doing so, it may become possible to establish in which scenarios correlations emerge as response-style artifacts and when they are more substantive. Future research is needed to test these ideas.
- **Designing user-centric products or services:** In the field of product design, personas are a widely used, yet criticized, method (Matthews et al., 2012; Miaskiewicz & Kozar, 2011). The use of LLM-generated personas could potentially speed up the design process and allow for product designs that better fit a diverse user base. Others suggest using personality dimensions to create non-player characters (NPCs) with human-like traits for use in computer games (Klinkert et al., 2024).

Apart from the above, we see future potential in applying ChatGPT to explore new methods of psychology research. Previous research has also shown that it is possible to emulate the cognitive skills of children of different ages using ChatGPT (Milička et al., 2024). Using personas that perform cognitive tasks could enable the testing of more sophisticated psychological theories, particularly those related to collaboration or joint intelligence.

4. Conclusion

The current findings represent only a first step toward the use of LLMs in personality research but are nonetheless promising. We have demonstrated that the Big Five personality dimensions can be identified through the use of virtual personas. Future research should focus on generating more realistic personas, and on working toward a prompting method that is aligned with human data. Ideally, this would allow for questionnaires to be pilot-tested before contacting actual humans. We anticipate that as better-quality LLMs emerge in the future, an increasing number of questions pertaining to personality, or psychology more broadly, can be proactively answered.

CRedit authorship contribution statement

Joost C.F. de Winter: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tom Driessen:** Conceptualization, Methodology, Writing – review & editing. **Dimitra Dodou:** Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The used personas, MATLAB scripts that reproduce the tables and figure, and the GPT output are downloadable from the following data repository: <https://doi.org/10.4121/6e0f2f2b-f1fc-4300-b8ca-eb9031a7b257>.

Appendix A. Supplementary Material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.paid.2024.112729>.

References

- Aluja, A., García, L. F., Rossier, J., Ostendorf, F., Glicksohn, J., Oumar, B., ... Hansenne, M. (2022). Dark triad traits, social position, and personality: A cross-cultural study. *Journal of Cross-Cultural Psychology*, 53, 380–402. <https://doi.org/10.1177/00220221211072816>
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31, 337–351. <https://doi.org/10.1017/pan.2023.2>
- Arias, V. B., Jenaro, C., & Ponce, F. P. (2018). Testing the generality of the general factor of personality: An exploratory bifactor approach. *Personality and Individual Differences*, 129, 17–23. <https://doi.org/10.1016/j.paid.2018.02.042>
- Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which humans? *PsyArXiv*. <https://doi.org/10.31234/osf.io/5b26t>
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671–732.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2, 396–403. <https://doi.org/10.1111/j.1745-6916.2007.00051.x>
- Biderman, M. D., Nguyen, N. T., Cunningham, C. J. L., & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the Big Five. *Journal of Research in Personality*, 45, 417–429. <https://doi.org/10.1016/j.jrp.2011.05.001>
- Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*. <https://doi.org/10.1017/pan.2024.5>
- Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin*, 130, 887–919. <https://doi.org/10.1037/0033-2909.130.6.887>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. <https://doi.org/10.48550/arXiv.2303.12712>
- Cao, X., & Kosinski, M. (2024). Large language models know how the personality of public figures is perceived by the general public. *Scientific Reports*, 14, 6735. <https://doi.org/10.1038/s41598-024-57271-z>
- Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, 38, 476–506. <https://doi.org/10.1037/h0054116>
- Cross, C. P., Cyrenne, D.-L. M., & Brown, G. R. (2013). Sex differences in sensation-seeking: A meta-analysis. *Scientific Reports*, 3, 2486. <https://doi.org/10.1038/srep02486>
- Crysel, L. C., Crosier, B. S., & Webster, G. D. (2013). The Dark Triad and risk behavior. *Personality and Individual Differences*, 54, 35–40. <https://doi.org/10.1016/j.paid.2012.07.029>
- Dahlen, E. R., & White, R. P. (2006). The Big Five factors, sensation seeking, and driving anger in the prediction of unsafe driving. *Personality and Individual Differences*, 41, 903–915. <https://doi.org/10.1016/j.paid.2006.03.016>

- Davies, S. E., Connelly, B. S., Ones, D. S., & Birkland, A. S. (2015). The general factor of personality: The "Big One," a self-evaluative trait, or a methodological gnat that won't go away? *Personality and Individual Differences*, 81, 13–22. <https://doi.org/10.1016/j.paid.2015.01.006>
- De Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*. <https://doi.org/10.1007/s11192-024-04939-y>
- De Winter, J. C. F. (2023). Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-023-00372-z>
- De Winter, J. C. F., & Dodou, D. (2016). National correlates of self-reported traffic violations across 41 countries. *Personality and Individual Differences*, 98, 145–152. <https://doi.org/10.1016/j.paid.2016.03.091>
- Denissen, J. J. A., Geenen, R., Van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, 90, 152–157. <https://doi.org/10.1080/00223890701845229>
- DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, 91, 1138–1151. <https://doi.org/10.1037/0022-3514.91.6.1138>
- Eachus, P. (2004). Using the Brief Sensation Seeking Scale (BSSS) to predict holiday preferences. *Personality and Individual Differences*, 36, 141–153. [https://doi.org/10.1016/S0191-8869\(03\)00074-6](https://doi.org/10.1016/S0191-8869(03)00074-6)
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40, 35–42. <https://doi.org/10.1016/j.soec.2010.10.008>
- Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59, 1216–1229. <https://doi.org/10.1037/0022-3514.59.6.1216>
- Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, 380, 1108–1109. <https://doi.org/10.1126/science.ad11778>
- Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17, 65–70. <https://doi.org/10.1111/j.2044-8317.1964.tb00244.x>
- Hilliard, A., Munoz, C., Wu, Z., & Koshiyama, A. S. (2024). Eliciting Big Five personality traits in large language models: A textual analysis with classifier-driven approach. *arXiv*. <https://doi.org/10.48550/arXiv.2402.08341>
- Hoyle, R. H., Stephenson, M. T., Palmgreen, P., Lorch, E. P., & Donohew, R. L. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences*, 32, 401–414. [https://doi.org/10.1016/S0191-8869\(01\)00032-0](https://doi.org/10.1016/S0191-8869(01)00032-0)
- Ji, Y., Wu, W., Zheng, H., Hu, Y., Chen, X., & He, L. (2023). Is ChatGPT a good personality recognizer? A preliminary study. *arXiv*. <https://doi.org/10.48550/arXiv.2307.03952>
- John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, 2, 171–203. <https://doi.org/10.1002/per.2410020302>
- Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment*, 21, 28041. <https://doi.org/10.1177/1073191113514105>
- Just, C. (2011). A review of literature on the general factor of personality. *Personality and Individual Differences*, 50, 765–771. <https://doi.org/10.1016/j.paid.2011.01.008>
- Klinkert, L. J., Buongiorno, S., & Clark, C. (2024). Driving generative agents with their personality. *arXiv*. <https://doi.org/10.48550/arXiv.2402.14879>
- Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., & Zhou, X. (2024). Better zero-shot reasoning with role-play prompting. *arXiv*. <https://doi.org/10.48550/arXiv.2308.07702>
- Lovik, A., Verbeke, G., & Molenberghs, G. (2017). Evaluation of a very short test to measure the Big Five personality factors on a Flemish sample. *Journal of Psychological and Educational Research*, 25, 7–17.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99. <https://doi.org/10.1037/1082-989X.4.1.84>
- Matthews, T., Judge, T., & Whittaker, S. (2012). How do designers and user experience professionals actually perceive and use personas?. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems-CHI '12* (pp. 1219–1228), Austin, TX. <https://doi.org/10.1145/2207676.2208573>
- Matz, S. C., Gladstone, J. J., & Stillwell, D. (2016). Money buys happiness when spending fits our personality. *Psychological Science*, 27, 715–725. <https://doi.org/10.1177/0956797616635200>
- Miaskiewicz, T., & Kozar, K. A. (2011). Personas and user-centered design: How can personas benefit product design processes? *Design Studies*, 32, 417–430. <https://doi.org/10.1016/j.destud.2011.03.003>
- Milická, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., & Drobil, O. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLoS One*, 19, Article e0298522. <https://doi.org/10.1371/journal.pone.0298522>
- Moritz, D., & Roberts, J. E. (2018). Self-other agreement and metaperception accuracy across the Big Five: Examining the roles of depression and self-esteem. *Journal of Personality*, 86, 296–307. <https://doi.org/10.1111/jopy.12313>
- Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The malevolent side of human nature: A meta-analysis and critical review of the literature on the dark triad (narcissism, Machiavellianism, and psychopathy). *Perspectives on Psychological Science*, 12, 183–204. <https://doi.org/10.1177/1745691616666070>
- Olea, C., Tucker, H., Phelan, J., Pattison, C., Zhang, S., Lieb, M., ... White, J. (2024). Evaluating persona prompting for question answering tasks. In *Proceedings of the 10th international conference on artificial intelligence and soft computing*, Sydney, Australia.
- OpenAI. (2023). *GPT-4 technical report*. *arXiv*. <https://doi.org/10.48550/arXiv.2303.08774>
- Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*, San Francisco, CA. <https://doi.org/10.1145/3586183.3606763>
- Peters, H., & Matz, S. (2023). *Large Language Models can infer psychological dispositions of social media users*. *arXiv*. <https://doi.org/10.48550/arXiv.2309.08631>
- Phillips, J., & Robie, C. (2024). Can a computer outfake a human? *Personality and Individual Differences*, 217, Article 112434. <https://doi.org/10.1016/j.paid.2023.112434>
- Poropat, A. E. (2009). A meta-analysis of the Five-Factor Model of personality and academic performance. *Psychological Bulletin*, 135, 322–338. <https://doi.org/10.1037/a0014996>
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 203–212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- Rao, H., Leung, C., & Miao, C. (2023). Can ChatGPT assess human personalities? A general evaluation framework. In H. Bouamar, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1184–1194). Singapore: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.84>
- Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14, 403–414. https://doi.org/10.1207/s15327906mbr1404_2
- Rosenberg, J., & Egbert, N. (2011). Online impression management: Personality traits and concerns for secondary goals as predictors of self-presentation tactics on Facebook. *Journal of Computer-Mediated Communication*, 17, 1–18. <https://doi.org/10.1111/j.1083-6101.2011.01560.x>
- Rutkowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2024). The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024, Article 7115633. <https://doi.org/10.1155/2024/7115633>
- Saucier, G., & Goldberg, L. R. (1996). Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality*, 10, 61–77. [https://doi.org/10.1002/\(SICI\)1099-0984\(199603\)10:1<61::AID-PER246>3.0.CO;2-D](https://doi.org/10.1002/(SICI)1099-0984(199603)10:1<61::AID-PER246>3.0.CO;2-D)
- Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, 94, 168–182. <https://doi.org/10.1037/0022-3514.94.1.168>
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Mataric, M. (2023). *Personality traits in large language models*. *arXiv*. <https://doi.org/10.48550/arXiv.2307.00184>
- Spector, P. E., & Brannick, M. T. (2009). Common method variance or measurement bias? The problem and possible solutions. In D. Buchanan, & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 346–362). London: Sage.
- Spector, P. E., Liu, C., & Sanchez, J. I. (2015). Methodological and substantive issues in conducting multinational and cross-cultural research. *Annual Review of Organizational Psychology and Organizational Behavior*, 2, 101–131. <https://doi.org/10.1146/annurev-orgpsych-032414-111310>
- Tabone, W., & De Winter, J. C. F. (2023). Using ChatGPT for human-computer interaction research: A primer. *Royal Society Open Science*, 10, Article 231053. <https://doi.org/10.1098/rsos.231053>
- Tang, R., Zhang, X., Ma, X., Lin, J., & Ture, F. (2024). Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2310.07712>
- Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University Chicago Press.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859–883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Tucker, L. R. (1955). The objective definition of simple structure in linear factor analysis. *Psychometrika*, 20, 209–225. <https://doi.org/10.1007/BF02289018>
- Van der Linden, D., Te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, 44, 315–327. <https://doi.org/10.1016/j.jrp.2010.03.003>
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *arXiv*. <https://doi.org/10.48550/arXiv.2203.11171>
- Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., & Ji, H. (2024). Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv*. <https://doi.org/10.48550/arXiv.2307.05300>