Not Yet: Large Language Models Cannot Replace Human Respondents for Psychometric Research

Pengda Wang[1], Huiqi Zou[2], Zihan Yan[3], Feng Guo[4],

Tianjun Sun[1], Ziang Xiao[2], and Bo Zhang[5,6]

1 Department of Psychological Sciences, Rice University

2 Department of Computer Science, Johns Hopkins University

3 Cognitive Science and Language Processing Informatics, University of Illinois Urbana-Champaign

4 Department of Psychology, University of Tennessee at Chattanooga

5 School of Labor and Employment Relations, University of Illinois Urbana-Champaign

6 Department of Psychology, University of Illinois Urbana-Champaign

[This is a pre-print and has not yet been peer-reviewed.]

Author Note

We have no known conflicts of interest to disclose.

Correspondence concerning this paper should be addressed to Pengda Wang, Rice University, 6100 Main Street, Houston, 77005, United States. Email: pw32@rice.edu

Abstract

Multiple studies have claimed that artificial intelligence (AI), particularly large language models (LLMs), can simulate human-like responses on various psychological tasks such that AI may replace human respondents for social science studies. However, this claim may be premature because of limitations in the design and evaluation metrics of previous studies. The present study aimed to provide a comprehensive evaluation of this claim, focusing on LLMs, by comparing six types of LLM-generated responses and human responses to the Big Five Inventory-2 (BFI-2) and the HEXACO-100 personality inventory. While previous research has primarily highlighted similarities between LLM-generated responses and human responses at the broad personality domain level in terms of descriptive statistics (mean and standard deviation), we took a closer look by first comparing descriptive statistics at the item, facet, and domain levels. Then, we performed a comprehensive psychometric analysis (e.g., model fit, factor loadings, inter-factor correlations) of LLM-generated responses to examine the degree to which LLM-generated responses produced similar results as those produced by human responses. Our findings indicated that although LLMs perform well in replicating broad-level patterns, they fall short at the item level, where subtle human differences are more accurately captured, and significant psychometric challenges remain when using LLM-generated responses. Additionally, we explore the influence of social desirability on LLM-generated responses and apply logistic regression to differentiate between LLM and human responses. We emphasize the importance of rigorous validation and adherence to psychometric principles when using LLMs for psychological research.

Keywords: artificial intelligence, large language model, personality, psychometrics, survey methodology

Not Yet: Large Language Models Cannot Replace Human Respondents for Psychometric Research

Constructs such as personality, attitudes, interests, and motivation are the backbone of psychological and many other social science research. Accurate measurement of these constructs is the foundational first step toward the scientific study of human psychology. Up to now, researchers still mainly rely on self-report scales to capture these constructs. However, the development and validation of psychological scales involves multiple rounds of data collection and revisions. Further, when applying a scale developed in context A to context B, researchers are recommended to conduct additional validation studies in the target population to ensure its applicability. While there is no doubt these steps are crucial for the validity of the scale, it is also true that scale development is inevitably resource-consuming, placing significant time and financial burdens on researchers. If there can be a way to reduce the burden of scale development studies without sacrificing the quality of the resulting scale, that would be a game-changer for psychometric research. Artificial intelligence (AI) is a strong candidate.

In the past few years, a series of studies have found that generative AI, particularly large language models (LLMs), demonstrated a remarkable ability to produce human-like responses on various psychological tasks (e.g., Caron & Srivastava, 2022; Jiang et al., 2023; Lampinen et al., 2024; Pellert et al., 2024; P. Wang et al., 2024). This capability of LLMs to generate human-like responses has sparked growing considerable interest in considering LLMs as a substitute for human respondents among social scientists (Aher et al., 2022; Argyle et al., 2023; Hämäläinen et al., 2023). If it is truly the case that LLMs can replace human respondents, it will revolutionize the entire field of social sciences, including psychometric research.

However, many previous studies comparing LLMs-generated and human responses are limited in their designs and evaluation metrics. One of the major limitations is that most of these studies only focused on comparisons at the scale level, such as similarities in the distributional properties of scale scores (e.g., means (*M*), standard deviations (*SD*)) and correlations among them. Meaningful differences at the item level are likely to be obscured by aggregation. For instance, if LLMs consistently underreport on one item and overreport on another item of the same construct compared to human respondents, the mean scores for LLMs and human respondents will be highly similar, despite the discrepancies at the item level. Such nuanced differences can only be revealed when we focus on item-level analysis. As scale development is inherently an iterative process aimed at identifying good items and modifying or removing poor ones, only through item-level analysis can we more directly address the question: "Can LLMs replace human respondents in psychometric research?"

Given that, the present study aims to provide a comprehensive investigation into the possibility of using LLMs to replace human respondents in psychometric research. We chose two personality inventories – the evaluatively-oriented Big Five Inventory-2 (BFI-2; Soto & John, 2017) based on the five-factor model of personality and the behaviorally-oriented HEXACO-100 Personality Inventory (Lee & Ashton, 2018) based on the six-factor model of personality – as the tool for data collection. Personality was chosen as the example topic domain because a large amount of rigorous and influential psychometric research has been conducted within this area, and personality assessment itself has been a key and active research direction. Six sets of LLMs responses to the BFI–2 and the HEXACO-100 were generated under different prompts and LLMs, which were then compared to human responses to the same questionnaires. The use of two personality inventories under different theoretical frameworks and generating multiple datasets under different prompts and LLMs allowed us to be more comprehensive and systematic in our investigation. Aside from comparing means, standard deviations, and Cronbach's alphas at the broad personality factor and facet levels as in many previous studies, we also performed more refined psychometric analyses at the item level, including detailed confirmatory factor analyses, exploration of how item social desirability may explain LLMs-human discrepancy, and to what degree LLMs-generated and human responses can be differentiated by the most powerful machine learning model. Through these efforts, we aim to provide a more direct and stronger answer to the question: "Can LLMs replace human respondents in psychometric research?"

How to Evaluate the Similarity between Human and LLM-Generated Responses?

To assess whether LLM-generated responses can replace human responses, we first need to decide on metrics to quantify their similarities. Below are the metrics we deem necessary if researchers plan to use LLM-generated responses for psychometric properties.

At the scale level, human and LLM scores should first show similar distributions. That is to say, the means and variances of scale scores in human responses should be similar to those in the LLM-generated responses. Reliability estimates based on human and LLM-generated responses should also be similar to ensure that LLMs responded to items with a similar degree of internal consistency as human respondents. Additionally, correlations between scores on the focal scale and scores on other scales can be used as another metric to assess the similarity between human and LLM-generated responses. If such correlations are similar between humans and LLMs, researchers can be sure that LLMs accurately capture the nomological network of the focal construct, an essential component of construct validity.

However, only focusing on scale-level metrics is insufficient. Psychometric research in essence is about selecting good items and dropping items. The quality of a scale depends on the quality of its constituent items. Therefore, to use LLM-generated responses to perform item selection, item-level metrics are of critical importance. When developing scales, researchers often aim for items that can adequately reflect individual differences (sufficient variability) and discriminate individuals high on the focal trait from those who are low on the focal trait (high factor loadings). Besides, researchers also aim to ensure that the covariance among the chosen set of items can be well approximated by the hypothesized factor structure (good model fit). Therefore, in the scale development process, researchers often focus on item variances, factor loadings, and the fit of the hypothesized model. If we want to replace human responses with LLM-generated responses for psychometric research, it is critical that LLM-generated responses can produce similar variance, factor loadings, and model fit as human responses.

LLMs and Prompting Methods

Since we want to use LLMs to replace human respondents, an important question is: What are LLMs? LLMs are advanced neural networks designed to model and generate human language. These models are built using deep learning techniques and are trained on extensive text datasets, enabling them to recognize linguistic patterns and generate coherent, contextually appropriate text. They can perform a wide range of tasks, such as text generation, translation, summarization, and question-answering. Notable examples of LLMs include OpenAI's GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023), and Meta's LLaMA (Touvron et al., 2023). Through the analysis and learning of large datasets, LLMs can recognize complex language patterns (based on extensive statistical analysis of data rather than real experiences) and, in some cases, demonstrate human-like reasoning and decision-making abilities (Brown et al., 2020; OpenAI, 2023).

Aside from the LLMs, how we prompt these models to carry out the intended tasks also matters. Prompting methods are techniques used to interact with LLMs by crafting specific instructions or questions, known as prompts, to guide the model's output. Effective prompting helps extract desired information or achieve specific results, enhancing the model's performance in tasks like answering questions, creative writing, or problem-solving. By refining prompts, users can leverage the full potential of LLMs to meet various needs. We adopt two commonly used prompting methods to instruct the LLMs to respond to the personality inventory in related research—*persona* and *shape* (see Method section).

Existing Work on Using LLMs to Respond to Personality Inventory

As we know, personality encodes rich and complex information in language and text (Goldberg, 1990; Saucier & Goldberg, 2001). So, LLMs are able to capture and model these encodings by learning from vast amounts of training data. For example, by providing different types of contexts (personality descriptions or diagnostic questions about personality traits), Caron and Srivastava (2022) demonstrated that LLMs can recognize the descriptions of personality traits exhibited in these contexts.

So a considerable amount of research has investigated the capability of LLMs to substitute humans in answering personality questionnaires. For example, Huang, Wang, Lam et al. (2023) and Huang, Wang, Li et al. (2023) found that LLMs, such as GPT-3.5, GPT-4, and LLaMA2, have the potential to simulate different personality traits and represent various groups when responding to scales. Jiang et al. (2023) examined LLMs by setting specific levels in five domains to answer the 44-item BFI scale, discovering that the self-reported personality scores of the LLMs were highly consistent with their specified personality settings. Serapio-García et al.

(2023) also identified similarities between the outputs of the PaLM model under prompt configurations to simulate human profiles and human personality measurement results.

However, much of the work on using LLMs to simulate human personality emphasizes their similarity to humans in the main domains of personality without validation from the lower-order structure and a psychometric perspective. This lack of such more detailed psychometric evidence leaves a significant gap in our understanding of how well these models can truly replicate the nuances of human personality. Domains, or traits, are the broad dimensions of personality; the subtleties within these traits (e.g., facet, item) are also important for accurate personality simulation.

Petrov et al. (2024) conducted a more comprehensive examination, including an exploration of the internal consistency and construct validity of LLM-generated responses. However, their study only focused on the GPT series' closed-source models and a single personality scale, the Big Five Inventory (BFI; John et al., 1991). Different personality scales exhibit differences in item design and structure. For example, the BFI-2 tends to include more evaluative items, while the HEXACO-100 focuses more on behavioral items. These design differences may affect the performance of LLMs when processing these scales. Without a detailed examination of these differences, our understanding of LLMs' capabilities remains incomplete. Evaluative items often require a nuanced understanding of context, while behavioral items may necessitate a model's ability to infer actions from personality descriptors. Furthermore, testing more scales enhances the external validity of the study. By evaluating the simulation performance of the LLM across different scales, we can gain a better understanding of the applicability and reliability of the conclusions.

Considering these concerns, this paper aims to build on previous research by comprehensively examining and evaluating the capabilities and limitations of LLMs in replacing human respondents for answering personality inventory from a psychometric perspective. We seek to answer the following core research question: Can LLMs replace human respondents in psychometric research?

Method

Measures

Big Five Inventory-2 (BFI-2; Soto & John, 2017): The BFI-2 was designed to capture three core facets of each of the Big Five personality factors: Open-Mindedness, Conscientiousness, Extraversion, Agreeableness, and Negative Emotionality (Neuroticism). Each facet was measured by two positively worded items and two negatively worded items, resulting in 60 items in total. Human respondents and LLMs were instructed to indicate the degree to which they agree with each item on a 5-point scale (1 = “Strongly disagree”, 2 = “Somewhat disagree”, 3 = “Neither agree nor disagree”, 4 = “Somewhat agree”, 5 = “Strongly agree”).

HEXACO-100 (Lee & Ashton, 2018): The HEXACO-100 was designed to measure four facets of each of the Big Six personality factors: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. Each facet was measured by four items. The major difference between the Big Five factors and the Big Six factors is the addition of the Honesty-Humility domain. The HEXACO-100 also included an additional interstitial facet Altruism that is expected to simultaneously load on Honesty–Humility, Emotionality, and Agreeableness. Each facet was measured by four 4 items, resulting in 100 items in total. Human respondents and LLMs were instructed to indicate the degree to which they agree with each item on a 5-point scale (1 = “Strongly disagree”, 2 = “Disagree”, 3 = “Neutral (neither agree nor disagree)”, 4 = “Agree”, and 5 = “Strongly agree).

Human Participants

In the present study, human responses to the BFI-2 were collected as a part of a broader project related to personality assessment from the Prolific. Respondents were instructed to respond to a set of demographic questions, a forced-choice measure of personality, the BFI-2, and a set of criterion measures. Respondents were compensated $3.75 for participation. In total, 1,559 respondents provided valid responses. On average, participants were in their early 40s (*M* = 42.29, *SD* = 11.79) with 49.20% women.

Human responses to the HEXACO were obtained from the Anglim et al. (2022). The data was collected from multiple sources, including college students, students taking MOOC courses, and fire fighters. More details on the data collection can be found in Anglim et al. (2022). In the present study, we used 7,204 valid observations. On average, participants were in their early 30s (*M* = 30.87, *SD* = 8.13) with 63.40% women.

Generating Responses with LLMs

To evaluate whether LLMs could substitute human participants, we investigate three LLMs, encompassing both closed-source and open-source models: GPT-3.5 (gpt-3.5-turbo-0125) (Ouyang et al., 2022), GPT-4 (gpt-4-0613) (OpenAI, 2023), and LLaMA 3 (llama-3-70b-instruct) (Meta-Llama, 2024). We include these models to allow for cross-validation and comparative analysis, ensuring that the findings are robust and not specific to a single model. Additionally, GPT-4 and LLaMA 3 are among the current leading models, making them ideal for assessing the latest advancements in LLM performance. Meanwhile, GPT-3.5 is a widely recognized and commonly used model, serving as a familiar baseline.

To ensure reproducibility, the temperature[1] of the LLMs was set to 0 to generate deterministic responses. And in the initial exploratory experiment, we found that a simple prompting method—such as providing demographics (i.e., race, age, gender), job titles they are applying for, and celebrity names—resulted in a lack of diversity in responses from most LLMs. Despite being prompted to simulate individuals demonstrating different personality traits, the LLMs often generated similar answers to various items. This phenomenon is consistent with the findings of Serapio-García et al. (2023). To generate simulation data effectively, it is important to provide the LLMs with diverse contextual information. This context helps the LLMs understand the assigned task and generate appropriate responses that align with the goals of the study, the application scenarios, and the specific behaviors or characteristics required. In this paper, we will mainly introduce two commonly used prompting methods in related research—*persona* and *shape*.

Persona

The *persona* method is based on the Persona-Chat dataset constructed by Zhang et al. (2018). The dataset consists of persona descriptions; each made up of five short sentences containing demographic information, collected through Amazon Mechanical Turk crowdsourcing. These persona descriptions were required to be rewritten to avoid sentence similarity or repetition (e.g., “I am very shy.” changing to “I am not a social person.”). Zhang et al. (2018) have demonstrated through machine learning model validation and human evaluations that such persona descriptions provide an effective method to enhance the level of personalization. Currently, incorporating personal profiles into prompts is widely used in research related to LLM agents (Park et al., 2023; X. Wang et al., 2023; Xi et al., 2023). We treat

[1] Temperature is a parameter in language models that controls the randomness of predictions, where lower values make the output more deterministic.

each persona description as an individual entity (i.e., a single subject) and randomly select 300 persona descriptions from the dataset as test samples. One example is, “I wear a lot of leather. I have boots I always wear. I sleep in late during the day. I listen to metal music. I have black spiky hair.”

Shape

The *shape* method is based on the work of Serapio-García et al. (2023), who introduced a prompting approach to shape synthetic personality in LLMs along desired dimensions. The researchers expanded upon Goldberg’s (1990) lexical hypothesis, expanding his list of 70 bipolar adjectives (Goldberg, 1992) to include 104 trait adjectives. Additionally, they employed linguistic qualifiers commonly used in Likert-type scales (Likert, 1932), such as “a bit,” “very,” and “extremely,” to set target levels for each adjective. This resulted in a fine-grained prompting method with nine levels: 1. extremely {low adjective}; 2. very {low adjective}; 3. {low adjective}; 4. a bit {low adjective}; 5. neither {low adjective} nor {high adjective}; 6. a bit {high adjective}; 7. {high adjective}; 8. very {high adjective}; 9. extremely {high adjective}. In our study, each prompt involves five randomly selected adjective markers from a specific personality domain. These markers are positioned after a consistent linguistic qualifier to set the prompt at one of nine intensity levels. For example, one prompt is: “You are extremely friendly, extremely energetic, extremely assertive, extremely bold, and extremely active.” We also randomly select 300 prompts here.

Evaluation Metrics

To quantify the degree of similarity between human and LLM-generated responses, we leveraged multiple metrics. At the domain score and facet score levels, we compared the mean, standard deviation, Cronbach’s alpha, and correlations among the scores. At the item level, we also compared the mean and standard deviation of item scores. Specifically, we used Mean Absolute Error (*MAE*) and Pearson correlation to quantify the degree of similarities on these metrics. Smaller *MAE* and higher Pearson correlations indicate higher similarity. Aside from these metrics, we additionally fitted a three-facet confirmatory factor analysis (CFA) model to responses to each domain of the BFI-2 and a four-facet CFA model to responses to each domain of the HEXACO-100. Aside from the facet structure of each domain, we also used the facet scores as indicators and fitted a five-factor model for the BFI-2 and a six-factor model for the HEXACO-100. Model fit, standardized factor loadings, and latent correlations among the facets were also compared between human and LLM-generated responses. Tucker's congruence coefficient (*TCC*) was used to quantify the degree to which the factor solution obtained from human responses is similar to that from LLM-generated responses.

Results: BFI-2 Explorations

Descriptive Statistics

We compared the means and standard deviations of the human responses and LLM-generated responses at three levels: item, facet, and domain. To assess the similarity between the two datasets, we calculated the *MAE* and Profile Correlation for both the means and standard deviations, as detailed in Table 1. The *MAE* for the means reflected the average difference between the LLM-generated responses and the human responses at each level, while the *MAE* for the standard deviations revealed the difference in variability between the two datasets. Profile correlation further illustrated the linear relationship between the two datasets, with values closer to 1 indicating a stronger correlation. Detailed means and standard deviations for human responses and LLM-generated responses at the item, facet, and domain levels can be found in Tables 23, 24, 25, 26, 27, and 28 in Appendix A.

We use the criteria of a mean *MAE* less than 0.5, a standard deviation *MAE* less than 0.3, and both profile correlations greater than .3 to determine if the data is close to human responses. It was observed that the mean values of LLM-generated responses were relatively close to those of the human responses, particularly at the domain level, where both the *MAE* and profile correlations indicated high similarity. For standard deviation, the GPT series models generated simulated data with high *MAE* differences when using the *persona* method, while the *shape* method resulted in smaller *MAE* values. The LLaMA3 model showed relatively low *MAE* values regardless of whether the persona or shape method was used. However, the profile correlations for standard deviation were either not correlated or negatively correlated in all cases.

Furthermore, there was a clear pattern for the mean values: the *MAE* decreased, and the profile correlation increased with higher aggregation levels (*MAE* also decreased for most standard deviations with higher aggregation levels). At the domain level, LLM-generated responses and human responses were quite similar, which aligned with previous research findings (Ai et al., 2024; Jiang et al., 2023; Serapio-García et al., 2023). However, our exploration at the item and facet levels indicated that this conclusion lacked detailed analysis at these stages. The aggregation of scores from item to facet to domain levels reduced the impact of extreme values, which were smoothed out at higher levels. Significant differences still existed at some item and facet levels in the LLM-generated responses (also see Appendix A, Tables 23, 24, 25, and 26).

Table 1

Mean Absolute Error and Profile Correlation for BFI-2 Human Responses and LLM-Generated Responses at Item, Facet, and Domain Levels

Level		LLM-Generated Responses											
		persona GPT3.5		*shape* GPT3.5		*persona* GPT4		*shape* GPT4		*persona* LLaMA3		*shape* LLaMA3	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Item level	*MAE*	0.33	0.55	0.45	0.25	0.53	0.73	0.49	0.24	0.51	0.26	0.62	0.25
	Profile Correlation	.81	-.24	.73	-.18	.60	-.25	.84	-.54	.57	.02	.45	-.26
Facet level	*MAE*	0.25	0.53	0.40	0.17	0.51	0.60	0.45	0.20	0.33	0.15	0.52	0.29
	Profile Correlation	.89	.05	.84	-.33	.77	-.51	.96	-.74	.80	.31	.78	-.59
Domain level	*MAE*	0.20	0.46	0.30	0.11	0.48	0.51	0.39	0.19	0.31	0.06	0.45	0.31
	Profile Correlation	.91	.19	.87	-.62	.91	-.92	.97	-.82	.80	.76	.82	-.65

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Psychometric Performance

Model Fit

For both human responses and LLM-generated responses, the Three-Factor model was fitted to each BFI-2 domain, and the Five-Factor model was fitted to all the data. Model fit information is shown in Table 2 and Table 3.

For the Three-Factor model of each domain, all fit indices suggested that the models fitted better to human responses than LLM-generated responses. For the Five-Factor model, the simulated data based on GPT series models and the *persona* method is relatively close to the human responses, while the other simulated data perform worse in model fitting compared to the human responses.

Table 2

Model Fits for BFI-2 Three-Factor Models of Each Domain

		Chi-square	*df*	CFI	TLI	RMSEA	SRMR
E	human responses	993.189	51	.892	.861	.109	.057
	persona GPT3.5	267.890	51	.764	.694	.119	.084
	shape GPT3.5	739.044	51	.749	.676	.212	.110
	persona GPT4	379.893	51	.805	.748	.147	.099
	shape GPT4	794.074	51	.798	.739	.220	.077
	persona LLaMA3	316.715	51	.885	.851	.132	.060
	shape LLaMA3	934.002	51	.684	.592	.241	.132
A	human responses	904.640	51	.875	.838	.104	.060
	persona GPT3.5	315.506	51	.692	.601	.132	.119
	shape GPT3.5	493.975	51	.844	.798	.170	.078
	persona GPT4	432.446	51	.769	.701	.158	.103
	shape GPT4	685.567	51	.884	.850	.204	.050
	persona LLaMA3	616.199	51	.798	.739	.192	.088
	shape LLaMA3	874.914	51	.840	.793	.233	.070

C	human responses	1041.784	51	.897	.867	.112	.058
	persona GPT3.5	268.434	51	.829	.778	.119	.100
	shape GPT3.5	785.307	51	.705	.619	.219	.139
	persona GPT4	556.734	51	.747	.673	.182	.116
	shape GPT4	907.749	51	.784	.721	.237	.090
	persona LLaMA3	705.359	51	.768	.699	.207	.109
	shape LLaMA3	1502.901	51	.642	.536	.310	.196
N	human responses	929.871	51	.931	.911	.105	.052
	persona GPT3.5	367.472	51	.661	.561	.144	.118
	shape GPT3.5	868.704	51	.618	.506	.231	.144
	persona GPT4	521.364	51	.754	.682	.175	.131
	shape GPT4	647.972	51	.813	.759	.198	.080
	persona LLaMA3	346.207	51	.884	.850	.139	.059
	shape LLaMA3	808.305	51	.772	.705	.224	.101
O	human responses	909.210	51	.899	.870	.104	.064
	persona GPT3.5	115.216	51	.880	.845	.065	.068
	shape GPT3.5	326.574	51	.862	.822	.134	.078
	persona GPT4	264.997	51	.847	.801	.118	.082
	shape GPT4	825.420	51	.812	.756	.225	.069
	persona LLaMA3	256.032	51	.904	.876	.116	.077
	shape LLaMA3	569.466	51	.861	.820	.185	.057

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

Table 3

Model Fits for BFI-2 Five-Factor Model

	Chi-square	*df*	CFI	TLI	RMSEA	SRMR
human responses	1747.433	80	.852	.806	.116	.080
persona GPT3.5	330.287	80	.861	.817	.102	.074
shape GPT3.5	1421.730	80	.734	.651	.236	.175
persona GPT4	601.517	80	.768	.695	.147	.097
shape GPT4	1858.255	80	.736	.654	.272	.116
persona LLaMA3	954.848	80	.780	.711	.191	.120
shape LLaMA3	1988.584	80	.717	.629	.283	.149

Note. n = 1559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Structural Validity

Tucker's congruence coefficient (*TCC*) was used to evaluate the similarity of factor loadings between human responses and LLM-generated responses. *TCC* results for the Three-Factor models of each BFI-2 domain and the Five-Factor model are shown in Table 4 and Table 5. Specific standardized factor loading results are shown in Appendix A, Tables 29 and 30.

As Lorenzo-Seva and Berge (2006) stated, a *TCC* above .95 indicates good similarity, while a *TCC* of .85 to .94 suggests fair similarity. For BFI-2 Three-Factor models of each domain, the factor structure of the simulated data generated by the GPT series models using the *persona* method still shows differences compared to human responses. This was especially noticeable in GPT-3.5, where there was even a case of a negative *TCC* (TCC_{Trust} = -.84). The data generated using the *shape* method generally performed better, with no significant differences across different models. Among all models, LLaMA3 performed the best; regardless of whether the simulated data were generated by the *persona* or *shape* method, their *TCCs* were above .95.

For the BFI-2 five-factor model, LLM-generated responses generally performed very well, with all *TCCs* above .95.

Although most *TCCs* were relatively good, some differences remained when examining the specific standardized factor loadings between the LLM-generated responses and the human responses (see Appendix Tables 29 and 30).

Table 4

Tucker's Congruence Coefficient for BFI-2 Three-Factor Models of Each Domain

		persona GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sociability	.99	.99	1.00	1.00	.99	.99
	Assertiveness	*.85*	*.93*	.97	.99	.98	.97
E	Energy Level	.98	.97	.97	.98	.98	.96
	Compassion	**.78**	.97	.96	.97	.96	.95
	Respectfulness	**.78**	1.00	.98	1.00	1.00	1.00
A	Trust	**-.84**	1.00	*.93*	1.00	.99	.99
	Organization	*.87*	.98	.98	.99	1.00	.95
	Productiveness	.97	.97	.96	.99	.99	.99
C	Responsibility	*.94*	.98	.95	.98	.98	.96
	Anxiety	.98	.98	.99	.98	.99	.99
	Depression	.88	.97	.98	1.00	.99	.99
N	Emotional Volatility	.97	.99	*.89*	1.00	1.00	.99
	Intellectual Curiosity	**.68**	.99	.97	1.00	1.00	.98
	Aesthetic Sensitivity	*.87*	1.00	.97	1.00	.98	.98
O	Creative Imagination	**.81**	.97	1.00	.99	.98	1.00

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Italics for *TCC* .85 to .949, and boldface for *TCC* lower than .85.

Table 5

Tucker's Congruence Coefficient for BFI-2 Five-Factor Model

	persona GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
E	1.00	1.00	.98	1.00	.99	1.00
A	.99	.99	.99	.99	1.00	.99
C	.99	.96	1.00	.99	.98	.99
N	.98	.99	.99	.99	1.00	1.00
O	1.00	.99	.96	1.00	.99	.99

Note. n = 1559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Italics for *TCC* .85 to .949, and boldface for *TCC* lower than .85.

We also present the inter-factor correlations for the BFI-2 Three-Factor models of each domain and the BFI-2 Five-Factor model in Table 6 and Table 7. We observed that there were differences in inter-factor correlations between LLM-generated responses and human responses. For BFI-2 Three-Factor models of each domain, the LLM-generated responses generated using the GPT-3.5 model and *persona* method performed the worst, with many inter-factor correlations differing by more than .20 compared to the human responses. Other LLM-generated responses generated using the *persona* method also showed differences from the human responses. Another clear pattern was that LLM-generated responses generated using the *shape* method generally exhibited larger inter-factor correlations compared to human responses. This indicated that LLM-generated responses generated using the shape method did not adequately capture the distinct yet correlated latent constructs within each domain, often treating them as homogeneous.

For the BFI-2 Five-Factor model, there were also obvious differences in inter-factor correlations between LLM-generated responses and human responses.

Table 6

Inter-factor Correlations for BFI-2 Three-Factor Models of Each Domain

		human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sociability~~Assertiveness	.59	*.76*	**1.09**	.66	**.84**	**.89**	**.91**
	Sociability~~ Energy Level	.64	.70	**.95**	**.30**	**.94**	**.84**	*.80*
E	Assertiveness~~Energy Level	.47	*.66*	**1.10**	.42	**.87**	**.82**	*.61*
	Compassion~~Respectfulness	.81	.87	*.98*	.84	*.92*	*.98*	.87
	Compassion~~Trust	.70	**-.87**	**1.03**	**.42**	**.96**	**.94**	**.96**
A	Respectfulness~~Trust	.58	**-.74**	**.94**	*.42*	**.93**	**.96**	**.84**
	Organization~~Productiveness	.75	.78	.76	.72	*.91*	.81	.70
	Organization~~Responsibility	.70	*.86*	**1.03**	.77	**.92**	*.85*	*.86*
C	Productiveness~~Responsibility	.86	.92	*.75*	.92	.86	.87	.89
	Anxiety~~Depression	.81	**1.03**	**1.12**	**.55**	*.95*	.88	*.96*
	Anxiety~~Emotional Volatility	.90	*.79*	.96	**.59**	.90	*.79*	.87
N	Depression~~Emotional Volatility	.79	.77	*.92*	*.63*	.85	.85	.83
	Intellectual Curiosity~~Aesthetic Sensitivity	.66	.69	**1.00**	**.42**	**.88**	.59	**1.00**
	Intellectual Curiosity~~Creative Imagination	.74	*.90*	.83	*.85*	**.97**	.67	*.89*
O	Aesthetic Sensitivity~~Creative Imagination	.63	*.49*	**.88**	.64	**.83**	.60	**.86**

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Table 7

Inter-factor Correlations for BFI-2 Five-Factor Model

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
E~~A	.28	.26	**.50**	**.52**	**.52**	*.10*	**.54**
E~~C	.56	.48	**.10**	*.45*	.53	**.26**	**.31**
E~~N	-.61	*-.74*	*-.78*	*-.48*	**-.82**	-.59	*-.72*
E~~O	.35	.39	**.67**	**.60**	**.70**	*.51*	**.71**
A~~C	.47	**.69**	*.57*	**.76**	**.76**	**.77**	*.65*
A~~N	-.43	*-.55*	**-.63**	*-.55*	**-.68**	*-.54*	**-.73**
A~~O	.27	*.38*	**.50**	*.43*	**.55**	.25	**.89**
C~~N	-.61	-.65	**-.41**	*-.47*	-.65	-.63	**-.82**
C~~O	.17	.26	**-.03**	*.28*	**.39**	.16	**.45**
N~~O	-.17	**-.37**	**-.38**	*-.28*	**-.53**	-.23	**-.65**

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Scale Reliability

Facet-level and domain-level Cronbach's alpha for LLM-generated responses on BFI-2 and human responses are shown in Table 8. There was an obvious difference in Cronbach's alpha between LLM-generated responses and human responses.

At the facet level, LLM-generated responses generated using the GPT-3.5 model and *persona* method performed the worst compared to human responses; most differences in Cronbach's alpha exceeded .20, and most Cronbach's alpha values were below .70. For LLM-generated responses generated by other model and method combinations, the Cronbach's alpha values were closer to those of the human responses, but many differences still exceeded .10. At the domain level, the LLM-generated responses and human responses were relatively close, but for data generated based on the GPT-3.5 model and *persona* method, most differences in Cronbach's alpha still exceeded .10 compared to human responses.

Table 8

Cronbach's alpha for BFI-2 Human Responses and LLM-Generated Responses

		human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sociability	.87	*.73*	.84	.86	.90	.86	*.77*
	Assertiveness	.81	**.29**	*.64*	*.68*	.87	.80	.86
	Energy Level	.73	**.67**	*.90*	*.77*	*.90*	*.86*	*.87*
	Compassion	.67	**.48**	*.83*	*.78*	**.92**	**.87**	**.93**
	Respectfulness	.74	**.37**	*.86*	.76	*.93*	*.85*	*.93*
	Trust	.80	**.44**	.85	*.69*	*.97*	.79	*.94*
	Organization	.87	**.58**	*.70*	**.59**	.85	.80	.82
	Productiveness	.80	*.68*	.83	.77	*.92*	.85	*.90*
	Responsibility	.77	.69	.81	.78	*.90*	*.88*	.86
Facet	Anxiety	.84	**.58**	*.68*	.81	.79	.81	.79

	Depression	.86	**.57**	.79	*.74*	.90	.88	.91
	Emotional Volatility	.89	**.59**	*.73*	*.71*	.88	.89	.88
	Intellectual Curiosity	.75	**.23**	.77	.70	*.90*	*.85*	*.89*
	Aesthetic Sensitivity	.83	**.54**	.82	.81	.92	.83	*.93*
	Creative Imagination	.82	**.45**	.76	*.68*	.92	.82	.89
	E	.90	*.79*	.92	.83	.95	.93	.92
	A	.88	*.71*	.94	.84	.97	.94	.97
	C	.92	.85	.90	.88	.95	.93	.94
	N	.95	*.80*	.89	.86	.94	.93	.94
Domain	O	.86	*.67*	.91	.84	*.96*	.89	*.96*

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Discriminant Validity

The results for discriminant validity are shown in Table 9. The correlation patterns between human responses and LLM-generated responses were consistent in terms of their direction, but it was clear that Big Five factors were substantially more distinct from each other in LLM-generated responses ($M_{\text{human responses}}$ = .28; $M_{\textit{persona}\text{ GPT3.5}}$ = .35; $M_{\textit{shape}\text{ GPT3.5}}$ = .41; $M_{\textit{persona}\text{ GPT4}}$ = .36; $M_{\textit{shape}\text{ GPT4}}$ = .57; $M_{\textit{persona}\text{ LLaMA3}}$ = .35; $M_{\textit{shape}\text{ LLaMA3}}$ = .62). In the LLM-generated responses, the differences observed between the Big Five factors were greater with the *shape* method than with the *persona* method ($M_{\textit{persona}\text{ GPT3.5}}$ = .35 vs. $M_{\textit{shape}\text{ GPT3.5}}$ = .41; $M_{\textit{persona}\text{ GPT4}}$ = .36 vs. $M_{\textit{shape}\text{ GPT4}}$ = .57; $M_{\textit{persona}\text{ LLaMA3}}$ = .35 vs. $M_{\textit{shape}\text{ LLaMA3}}$ = .62).

Table 9

Domain Level Correlation Analysis for BFI-2 Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
E, A	.17	.18	*.31*	*.36*	**.44**	.12	**.40**
E, C	.35	.37	.35	.38	**.55**	.33	.41
E, N	-.45	-.48	*-.59*	*-.34*	*-.64*	-.42	*-.57*
E, O	.22	.23	**.62**	*.33*	**.66**	.28	**.68**
A, C	.33	**.55**	*.49*	**.63**	**.73**	**.66**	**.70**
A, N	-.36	-.44	*-.50*	-.41	**-.66**	*-.54*	**-.71**
A, O	.20	.27	*.33*	**.40**	**.55**	*.31*	**.78**
C, N	-.49	-.57	*-.63*	*-.37*	**-.71**	*-.60*	**-.87**
C, O	.09	*.20*	.05	*.26*	**.42**	*.20*	**.48**
N, O	-.11	*-.21*	*-.24*	-.17	**-.39**	-.08	**-.57**
Mean of absolute values	.28	.35	*.41*	.36	**.57**	.35	**.62**

Note. n = 1,559 for human responses, *n* = 299 for *persona* GPT-3.5, *n* = 297 for *shape* LLaMA-3, and *n* = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Results: HEXACO-100 Explorations

Descriptive Statistics

We presented the *MAE* and Profile Correlation for the means and standard deviations of both human responses and LLM-generated responses across item, facet, and domain levels in Table 10. For detailed means and standard deviations of the human responses and

LLM-generated responses at these levels, please refer to Tables 31, 32, 33, 34, 35, and 36 in Appendix A.

It can be seen that the performance of LLM-generated responses on HEXACO-100 was similar to that on BFI-2. They were closer in mean but had a large difference in standard deviation. Although the *MAE* of standard deviation is small, most of the profile correlations are weakly correlated, negatively correlated, or not correlated. The variation in means across different levels was similar to that in the BFI-2, primarily because extreme values are smoothed out at higher levels.

Table 10

Mean Absolute Error and Profile Correlation for HEXACO-100 Human Responses and LLM-Generated Responses at Item, Facet, and Domain Levels

Level		LLM-Generated Responses											
		persona GPT3.5		*shape* GPT3.5		*persona* GPT4		*shape* GPT4		*persona* LLaMA3		*shape* LLaMA3	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Item level	*MAE*	0.36	0.47	0.52	0.18	0.40	0.68	0.41	0.19	0.47	0.16	0.47	0.37
	Profile Correlation	.56	-.17	.38	-.06	.55	-.11	.54	-.36	.55	.24	.40	-.04
Facet level	*MAE*	0.23	0.40	0.34	0.13	0.36	0.49	0.31	0.19	0.34	0.09	0.39	0.41
	Profile Correlation	.71	-.22	.55	-.11	.75	-.09	.75	-.37	.65	.25	.46	-.07
Domain level	*MAE*	0.17	0.31	0.32	0.12	0.33	0.37	0.30	0.22	0.22	0.09	0.37	0.46
	Profile Correlation	.76	-.64	.38	.24	.80	.03	.75	.41	.52	-.13	.19	.09

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Psychometric Performance

Model Fit

For both human responses and LLM-generated responses, the Four-Factor model was fitted to each HEXACO-100 domain, and the Six-Factor model was fitted to all the data. The fit information for these models is presented in Tables 11 and 12.

Just like the responses generated by LLM for BFI-2, the simulated data generated by LLMs for HEXACO-100 also fit the models much less well compared to the human responses. Moreover, the Six-Factor model failed to converge for data generated by the combination of GPT-4 and the *persona* method, further highlighting the differences between LLM-generated responses and human responses.

Table 11

Model Fits for HEXACO-100 Four-Factor Models of Each Domain

		Chi-square	*df*	CFI	TLI	RMSEA	SRMR
	human responses	1753.428	98	.944	.931	.048	.041
	persona GPT3.5	250.082	98	.711	.646	.072	.079
	shape GPT3.5	422.692	98	.843	.808	.105	.082
	persona GPT4	603.754	98	.720	.658	.131	.113
	shape GPT4	1435.273	98	.728	.667	.213	.203
	persona LLaMA3	517.109	98	.870	.841	.119	.102
Hon	*shape* LLaMA3	971.395	98	.843	.808	.172	.097
	human responses	3383.079	98	.882	.856	.068	.056
	persona GPT3.5	300.170	98	.618	.532	.083	.087
	shape GPT3.5	1317.644	98	.627	.543	.204	.208
	persona GPT4	630.996	98	.619	.533	.135	.142
	shape GPT4	1306.996	98	.727	.666	.203	.219
	persona LLaMA3	785.698	98	.657	.581	.153	.137
Emo	*shape* LLaMA3	1204.774	98	.723	.661	.194	.185

	human responses	3256.230	98	.913	.894	.067	.048
	persona GPT3.5	348.806	98	.786	.738	.093	.086
	shape GPT3.5	744.899	98	.810	.767	.148	.082
	persona GPT4	510.420	98	.856	.824	.118	.087
	shape GPT4	1446.323	98	.779	.729	.214	.101
	persona LLaMA3	480.165	98	.851	.818	.114	.088
Ext	*shape* LLaMA3	1002.149	98	.790	.743	.175	.109
	human responses	2339.597	98	.923	.905	.056	.037
	persona GPT3.5	377.236	98	.618	.532	.098	.094
	shape GPT3.5	1216.776	98	.650	.571	.195	.139
	persona GPT4	870.552	98	.696	.628	.162	.165
	shape GPT4	1477.284	98	.746	.689	.217	.139
	persona LLaMA3	516.651	98	.800	.755	.119	.081
Agr	*shape* LLaMA3	1275.160	98	.723	.661	.200	.112
	human responses	3502.578	98	.875	.847	.069	.052
	persona GPT3.5	273.699	98	.798	.753	.078	.071
	shape GPT3.5	950.902	98	.598	.508	.170	.171
	persona GPT4	781.170	98	.695	.627	.152	.103
	shape GPT4	1388.463	98	.708	.643	.210	.127
	persona LLaMA3	666.770	98	.806	.763	.139	.101
Con	*shape* LLaMA3	1760.153	98	.620	.534	.238	.149
	human responses	2380.154	98	.900	.878	.057	.042
	persona GPT3.5	365.213	98	.602	.513	.096	.107
	shape GPT3.5	829.470	98	.748	.692	.158	.102
	persona GPT4	530.337	98	.683	.612	.121	.108
	shape GPT4	1440.581	98	.712	.647	.214	.139
	persona LLaMA3	390.484	98	.809	.766	.100	.096
Ope	*shape* LLaMA3	997.778	98	.807	.763	.175	.105

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience.

Table 12

Model Fits for HEXACO-100 Six-Factor Model

	Chi-square	*df*	CFI	TLI	RMSEA	SRMR
human responses	11873.011	237	.761	.721	.083	.077
persona GPT3.5	813.870	237	.658	.602	.090	.104
shape GPT3.5	2655.845	237	.659	.603	.184	.170
persona GPT4	NA	NA	NA	NA	NA	NA
shape GPT4	4727.102	237	.577	.507	.251	.240
persona LLaMA3	2385.234	237	.604	.539	.174	.171
shape LLaMA3	4872.257	237	.575	.505	.255	.227

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. The model did not converge for data generated by the combination of the GPT4 and *persona* methods.

Structural Validity

TCC results for the Four-Factor models of each HEXACO-100 domain and the Six-Factor model are shown in Table 13 and Table 14. Specific standardized factor loading results are shown in Appendix A, Tables 37 and 38.

Similarly, as with the performance of LLM-generated responses on BFI-2, for HEXACO-100 Four-Factor models of each domain, the factor structure of the simulated data generated by the GPT series models still showed differences from human responses, with many *TCCs* less than .85 or even negative. Among all the models, LLaMA3 performed the best, with the lowest *TCC* being .92. For the HEXACO-100 Six-Factor model, LLM-generated responses and human responses are similar in most factors except for Emotionality, where most of the *TCCs* were less than .85.

Table 13

Tucker's Congruence Coefficient for HEXACO-100 Four-Factor Models of Each Domain

		persona GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sincerity	**.44**	*.87*	*.92*	*.81*	*.94*	.98
	Fairness	.95	.99	.97	.99	1.00	.99
	Greed Avoidance	*.93*	.97	1.00	**-.85**	1.00	1.00
Hon	Modesty	*.93*	**-.75**	.99	*.87*	.98	.99
	Fearfulness	**-.46**	*.85*	*.92*	.99	.99	*.93*
	Anxiety	*.90*	*.90*	.99	.99	.99	.99
	Dependence	**.84**	.99	*.88*	*.86*	*.94*	1.00
Emo	Sentimentality	*.92*	.99	.95	.99	.95	.99
	Social Self-Esteem	**.60**	*.94*	.98	.98	.99	1.00
	Social Boldness	*.88*	.97	1.00	.99	1.00	.99
	Sociability	.99	1.00	.99	1.00	.99	1.00
Ext	Liveliness	**.24**	**.71**	1.00	1.00	.99	1.00
	Forgiveness	.98	.98	**.40**	1.00	.98	.97
	Gentleness	*.94*	.99	.96	.99	.98	*.94*
	Flexibility	*.88*	.99	.99	1.00	.99	.98
Agr	Patience	*.88*	**.72**	**.75**	.99	.97	.98
	Organization	.91	.99	*.88*	.96	*.94*	.97
	Diligence	*.89*	*.85*	.95	.99	.98	.98
	Perfectionism	*.93*	*.94*	.99	.99	.99	.97
Con	Prudence	.98	.97	.98	.97	.97	.96
	Aesthetic Appreciation	**.36**	.98	**.77**	.99	.99	.99
	Inquisitiveness	**-.53**	1.00	**.59**	.96	*.92*	*.93*
	Creativity	**.79**	.97	.96	.97	1.00	.98
Ope	Unconventionality	*.87*	*.94*	.95	*.93*	**.83**	.97

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion;

Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience. Italics for *TCC* .85 to .949, and boldface for *TCC* lower than .85.

Table 14

Tucker's Congruence Coefficient for HEXACO-100 Six-Factor Model

	persona GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Hon	.96	.99	NA	.99	.99	.99
Emo	**.79**	**.72**	NA	*.92*	**.77**	.98
Ext	.98	.98	NA	.99	1.00	.99
Agr	.99	.99	NA	1.00	1.00	1.00
Con	.99	.98	NA	.97	.99	.98
Ope	.96	.99	NA	.99	.98	.99

Note. *n* = 7,204 for human responses, *n* = 298 for *persona* GPT-3.5, and *n* = 300 for other LLM-generated responses. Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience. Italics for *TCC* .85 to .949, and boldface for *TCC* lower than .85.

Inter-factor correlations for the HEXACO-100 Four-Factor models of each domain and the HEXACO-100 Six-Factor model are shown in Table 15 and Table 16. Compared to the BFI-2, LLM-generated responses on HEXACO-100 showed greater differences from human responses. Regardless of the model or method used to generate the simulated data, there were differences in factor correlations exceeding .20 when compared to human responses. Additionally, a considerable number of factor correlations were negative or greater than 1. For the HEXACO-100 Six-Factor model, there were also similar differences in factor correlations between the LLM-generated responses and human responses.

Table 15

Inter-factor Correlations for HEXACO-100 Four-Factor Models of Each Domain

		human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sincerity~~Fairness	.50	*.40*	*.33*	**.78**	.56	**.89**	**.94**
	Sincerity~~Greed Avoidance	.42	.50	**.70**	.40	**-.84**	*.57*	**.93**
	Sincerity~~Modesty	.36	.33	**-.82**	*.54*	**.80**	**.59**	**.97**
	Fairness~~Greed Avoidance	.38	*.56*	**.63**	*.22*	**-.56**	.40	**.72**
	Fairness~~Modesty	.31	*.20*	**-.69**	.32	**.59**	*.42*	**.80**
Hon	Greed Avoidance~~Modesty	.61	.68	**-.91**	**.84**	**-.89**	**.83**	**.96**
	Fearfulness~~Anxiety	.57	**-.30**	**.84**	**.31**	**.89**	**.87**	**.86**
	Fearfulness~~Dependence	.42	**-.24**	**.07**	**.15**	**-.20**	.51	**.22**
	Fearfulness~~Sentimentality	.45	**-.09**	**-.28**	**.01**	**.09**	*.34*	**.19**
	Anxiety~~Dependence	.43	*.54*	*.53*	**-.02**	**.11**	.44	.49
	Anxiety~~Sentimentality	.53	*.35*	**.14**	**.13**	*.38*	.44	.52
Emo	Dependence~~Sentimentality	.57	**.83**	**1.03**	.58	**.95**	**.90**	**1.10**
	Social Self-Esteem~~Social Boldness	.56	.54	**.94**	**.27**	**.76**	*.69*	**.85**
	Social Self-Esteem~~Sociability	.48	.54	**.81**	*.33*	**.83**	*.64*	*.62*
	Social Self-Esteem~~Liveliness	.86	.95	*1.03*	.82	*.96*	.94	.92
	Social Boldness~~Sociability	.70	*.87*	**.90**	*.81*	.67	.76	.65
	Social Boldness~~Liveliness	.60	.59	**.98**	*.45*	**.93**	.64	**.98**
Ext	Sociability~~Liveliness	.60	.66	**.92**	*.43*	**.86**	.65	**.88**

	Forgiveness~~Gentleness	.58	*.68*	**.89**	**.90**	**1.00**	**.96**	**.81**
	Forgiveness~~Flexibility	.57	**1.14**	**.94**	**.12**	*.67*	**.91**	**.83**
	Forgiveness~~Patience	.61	**1.06**	*.79*	**.09**	*.77*	**.95**	*.73*
	Gentleness~~Flexibility	.73	.87	.79	.56	.72	.84	.91
	Gentleness~~Patience	.70	*.88*	.62	*.55*	*.80*	*.85*	.75
Agr	Flexibility~~Patience	.65	*.79*	*.83*	**.85**	**.90**	**1.01**	**.94**
	Organization~~Diligence	.60	.64	*.50*	.66	**.92**	.65	.65
	Organization~~Perfectionism	.51	**.73**	**.77**	**.87**	**.91**	**.84**	**.79**
	Organization~~Prudence	.62	*.73*	.68	**.91**	**.42**	**.89**	**.90**
	Diligence~~Perfectionism	.56	*.69*	*.43*	*.71*	**.84**	**.80**	**.78**
	Diligence~~Prudence	.54	.53	**-.06**	*.39*	**.09**	.49	*.42*
Con	Perfectionism~~Prudence	.50	**.84**	.48	**.76**	.48	**.80**	**.79**
	Aesthetic Appreciation~~Inquisitiveness	.60	**.09**	**1.12**	**.07**	**.98**	**.84**	**.99**
	Aesthetic Appreciation~~Creativity	.70	**-.38**	**.90**	.73	**.97**	.71	.79
	Aesthetic Appreciation~~Unconventionality	.68	**.06**	**1.09**	*.57*	**.91**	**.40**	**.98**
	Inquisitiveness~~Creativity	.38	**-.09**	**.94**	**.03**	**.89**	.38	**.71**
	Inquisitiveness~~Unconventionality	.57	**-.07**	**1.03**	**.34**	**.90**	**.32**	**.94**
Ope	Creativity~~Unconventionality	.68	*.83*	**1.05**	**.30**	**.92**	.63	**.92**

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Table 16

Inter-factor Correlations for the HEXACO-100 Six-Factor Model

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Hon~~Emo	-.21	**.16**	-.18	NA	**.37**	**.39**	**.17**
Hon~~Ext	.18	.25	.13	NA	**-.15**	**-.07**	.24
Hon~~Agr	.45	*.58*	**.73**	NA	*.59*	*.63*	**.83**
Hon~~Con	.42	.42	*.60*	NA	.43	*.31*	**.65**
Hon~~Ope	.04	**.27**	**.43**	NA	*.18*	**.26**	**.62**
Emo~~Ext	-.56	**-.04**	**-.90**	NA	**-.84**	**.18**	**-.77**
Emo~~Agr	-.51	**.07**	*-.63*	NA	**-.22**	**.43**	**-.21**
Emo~~Con	-.20	-.21	**-.62**	NA	*-.35*	**.29**	*-.37*
Emo~~Ope	.03	*.20*	**-.64**	NA	**-.20**	**.59**	**-.18**
Ext~~Agr	.43	*.61*	**.68**	NA	*.54*	.43	.52
Ext~~Con	.44	**.72**	**.64**	NA	*.63*	.53	*.58*
Ext~~Ope	.04	**.34**	**.76**	NA	**.57**	*.19*	**.73**
Agr~~Con	.26	**.56**	**.70**	NA	**.47**	*.43*	**.54**
Agr~~Ope	.00	*.19*	**.68**	NA	**.59**	**.51**	**.71**
Con~~Ope	.01	.00	**.57**	NA	**.44**	**.33**	**.43**

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. The model did not converge for data generated by the combination of the GPT4 and *persona* methods. Hon = Honesty-Humility; Emo = Emotionality;

Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Scale Reliability

Facet-level and domain-level Cronbach's alpha for LLM-generated responses on HEXACO-100 and human responses are shown in Table 17. Just like the BFI-2 results, there was a noticeable difference in Cronbach's alpha between the LLM-generated responses and the human responses.

Table 17

Cronbach's alpha for BFI-2 Human Responses and LLM-Generated Responses

		human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sincerity	0.65	**0.22**	0.56	*0.49*	0.64	0.69	*0.82*
	Fairness	0.75	**0.51**	0.76	0.83	*0.90*	*0.86*	*0.94*
	Greed Avoidance	0.79	**0.56**	0.77	0.76	*0.66*	0.88	*0.92*
	Modesty	0.68	**0.22**	**0.34**	*0.49*	0.59	*0.85*	*0.84*
	Fearfulness	0.69	**-0.02**	0.62	*0.59*	*0.82*	*0.59*	0.70
	Anxiety	0.72	**0.48**	0.68	0.80	*0.85*	0.73	*0.88*
	Dependence	0.73	**0.19**	*0.58*	**0.52**	0.65	*0.55*	0.68
	Sentimentality	0.68	*0.55*	*0.87*	0.60	**0.91**	0.75	**0.90**
	Social Self-Esteem	0.70	**0.38**	0.73	0.63	**0.91**	*0.87*	**0.91**
	Social Boldness	0.69	**0.41**	0.78	*0.87*	*0.85*	*0.80*	*0.82*
	Sociability	0.75	0.67	*0.88*	*0.88*	**0.96**	0.75	*0.92*
	Liveliness	0.79	**-0.16**	**0.46**	0.72	0.88	0.76	0.79
	Forgiveness	0.76	**0.40**	0.77	**-0.03**	*0.89*	*0.66*	0.85
	Gentleness	0.63	0.60	**0.83**	0.70	**0.87**	*0.76*	*0.81*
	Flexibility	0.61	**0.41**	**0.84**	**0.84**	**0.85**	*0.72*	**0.82**
	Patience	0.76	**0.34**	**0.38**	*0.59*	0.85	0.74	0.82
	Organization	0.74	**0.50**	0.69	**0.54**	0.76	0.72	0.78
	Diligence	0.69	*0.54*	0.61	0.67	**0.89**	*0.83*	*0.85*
	Perfectionism	0.67	*0.48*	0.66	0.76	**0.87**	*0.81*	*0.82*
	Prudence	0.67	*0.55*	0.75	0.73	*0.81*	0.85	*0.86*

Facet

	Aesthetic Appreciation	0.65	**0.09**	0.65	**0.35**	*0.82*	0.64	*0.84*
	Inquisitiveness	0.70	**0.31**	0.78	**0.35**	*0.80*	*0.57*	*0.81*
	Creativity	0.71	**0.50**	0.73	0.80	*0.86*	*0.81*	*0.85*
	Unconventionality	0.50	0.46	*0.68*	0.52	**0.83**	0.58	**0.84**
	Altruism	0.49	*0.31*	**0.79**	*0.59*	**0.94**	**0.72**	**0.91**
Domain	Hon	0.82	**0.61**	0.84	0.82	0.89	*0.92*	*0.96*
	Emo	0.82	**0.59**	0.81	*0.70*	0.89	0.83	0.89
	Ext	0.87	*0.68*	0.92	0.86	0.96	0.91	0.95
	Agr	0.85	*0.72*	0.91	0.78	*0.96*	0.90	0.94
	Con	0.83	0.78	0.85	0.86	0.92	0.92	0.92
	Ope	0.81	*0.65*	*0.92*	0.72	*0.95*	0.84	*0.95*

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Discriminant Validity

The results for discriminant validity are shown in Table 18. Compared to the BFI-2 result, the differences in discriminant validity between the human responses and LLM-generated responses were larger, with instances such as a negative correlation in the human responses turning positive in the LLM-generated responses (Hon, $\text{Emo}_{\text{human responses}}$ = -.12; Hon, $\text{Emo}_{\textit{persona}\ \text{GPT3.5}}$ = .05; Hon, $\text{Emo}_{\textit{shape}\ \text{GPT3.5}}$ = .28; Hon, $\text{Emo}_{\textit{persona}\ \text{GPT4}}$ = .20; Hon, $\text{Emo}_{\textit{shape}\ \text{GPT4}}$ = .30; Hon, $\text{Emo}_{\textit{persona}\ \text{LLaMA3}}$ = .28; Hon, $\text{Emo}_{\textit{shape}\ \text{LLaMA3}}$ = .29).

And similar to the findings in BFI-2, the differences between factors in the HEXACO-100 model were larger in the LLM-generated responses ($M_{\text{human responses}}$ = .16; $M_{\textit{persona}\ \text{GPT3.5}}$ = .21; $M_{\textit{shape}\ \text{GPT3.5}}$ = .40; $M_{\textit{persona}\ \text{GPT4}}$ = .19; $M_{\textit{shape}\ \text{GPT4}}$ = .37; $M_{\textit{persona}\ \text{LLaMA3}}$ = .26; $M_{\textit{shape}\ \text{LLaMA3}}$

= .42), and the differences observed using the *shape* method are greater than those observed using the *persona* method ($M_{persona\ \mathrm{GPT3.5}}$ = .21 vs. $M_{shape\ \mathrm{GPT3.5}}$ = .40; $M_{persona\ \mathrm{GPT4}}$ = .19 vs. $M_{shape\ \mathrm{GPT4}}$ = .37; $M_{persona\ \mathrm{LLaMA3}}$ = .26 vs. $M_{shape\ \mathrm{LLaMA3}}$ = .42).

Table 18

Domain Level Correlation Analysis for HEXACO-100 Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Hon~~Emo	-.12	*.05*	**.28**	**.20**	**.30**	**.28**	**.29**
Hon~~Ext	.08	.17	*.19*	**.28**	.04	*-.08*	*.23*
Hon~~Agr	.32	.30	**.71**	.31	**.63**	**.57**	**.77**
Hon~~Con	.28	.25	**.54**	**.53**	**.63**	.32	**.65**
Hon~~Ope	.04	*.22*	**.39**	**.29**	*.23*	.12	**.51**
Emo~~Ext	-.27	**-.05**	-.18	**-.07**	-.20	-.18	-.28
Emo~~Agr	-.33	**-.02**	**.03**	*-.19*	**.20**	**.14**	**.06**
Emo~~Con	-.13	-.13	-.11	*.02*	*-.01*	**.08**	-.17
Emo~~Ope	-.03	*.16*	*.16*	*.15*	**.35**	**.25**	**.24**
Ext~~Agr	.31	*.42*	**.57**	**.06**	**.55**	.33	*.48*
Ext~~Con	.25	**.51**	**.54**	.32	**.49**	*.44*	**.49**
Ext~~Ope	.06	*.23*	**.70**	.11	**.57**	*.19*	**.71**
Agr~~Con	.18	**.38**	**.53**	.22	**.48**	*.37*	**.46**
Agr~~Ope	.02	*.16*	**.57**	.03	**.53**	**.32**	**.60**
Con~~Ope	.02	.05	**.54**	.08	**.33**	*.18*	**.38**
Mean of absolute values	.16	.21	**.40**	.19	**.37**	*.26*	**.42**

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. Hon = Honesty Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience. Italics for absolute differences compared to the human responses of .1 to .199, and boldface for differences of .2 or higher.

Summary of BFI-2 and HEXACO-100 Results

Given the large volume of results presented above, we summarize the main findings in Table 19. Specifically, we compared LLM-generated responses and human responses on the BFI-2 and HEXACO-100. As shown in Table 19, there were notable differences between LLM-generated responses and human responses, regardless of the model and method used to generate the LLM-generated responses. The mean values in the descriptive statistics were the only dimension where the two datasets were similar. Other than that, both the standard deviation in the descriptive statistics and the psychometric performance, such as model fit and structural validity, were unsatisfactory.

The fact that the "≈" sign appears less frequently in Table 19 for the HEXACO-100 summary compared to the BFI-2 indicates that the LLM-generated responses on HEXACO-100 performs worse than the LLM-generated responses on BFI-2, particularly in the factor model fit and domain reliability. One possible reason for the poorer performance of LLM-generated responses on the HEXACO-100 compared to the BFI-2 may lie in the structural complexity and theoretical differences between the two personality models. The BFI-2 is based on the well-established Big Five personality framework, which has been extensively studied and widely applied across various domains. Consequently, LLMs may have had more exposure to data and patterns related to the Big Five traits, enabling them to generate responses more aligned with human behavior. In contrast, the HEXACO model introduces additional complexity by incorporating a sixth dimension, Honesty-Humility, which may be less familiar to LLMs due to its relatively recent emergence in personality research. The data used to train LLMs might not adequately cover the theoretical foundations of the HEXACO model.

Table 19

Results Summary for LLM-Generated Responses and Human Responses on the BFI-2 and HEXACO-100 Assessments

BFI-2								
Aspects	Dimensions		*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Descriptive statistics	Item	Mean	≈	≈	≠	≈	≠	≠
		Standard deviation	≠	≠	≠	≠	≠	≠
	Facet	Mean	≈	≈	≠	≈	≈	≠
		Standard deviation	≠	≠	≠	≠	≈	≠
	Domain	Mean	≈	≈	≈	≈	≈	≈
		Standard deviation	≠	≠	≠	≈	≈	≠
Psychometric performance	Model of each domain	Model fit	≠	≠	≠	≠	≠	≠
		Structural validity	≠	≠	≠	≠	≠	≠
	Five-Factor model	Model fit	≈	≠	≈	≠	≠	≠
		Structural validity	≠	≠	≠	≠	≠	≠
	Facet	Scale reliability	≠	≠	≠	≠	≠	≠
	Domain	Scale reliability	≠	≈	≈	≠	≈	≠
	Discriminant validity		≠	≠	≠	≠	≠	≠
HEXACO-100								
Aspects	Dimensions		*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Descriptive statistics	Item	Mean	≈	≠	≈	≈	≈	≈
		Standard deviation	≠	≠	≠	≠	≠	≠

		Mean	≈	≈	≈	≈	≈	≈
	Facet	Standard deviation	≠	≠	≠	≠	≠	≠
		Mean	≈	≈	≈	≈	≈	≠
	Domain	Standard deviation	≠	≠	≠	≈	≠	≠
	Model of each domain	Model fit	≠	≠	≠	≠	≠	≠
		Structural validity	≠	≠	≠	≠	≠	≠
		Model fit	≠	≠	NA	≠	≠	≠
	Six-Factor model	Structural validity	≠	≠	NA	≠	≠	≠
	Facet	Scale reliability	≠	≠	≠	≠	≠	≠
	Domain	Scale reliability	≠	≠	≠	≠	≠	≠
Psychometric performance	Discriminant validity		≠	≠	≠	≠	≠	≠

Note. For BFI-2, n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses; for HEXACO-100, n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. If the performance of LLM-generated responses was nearly equal to that of human responses, an "≈" sign was placed; if the performance differed from human responses, a "≠" sign was placed.

Additional Analyses: Social Desirability Rating and LLM Responses

From the findings above, the most noticeable point was the certain differences between LLM-generated responses and human responses. What could have caused this difference? A reasonable hypothesis is that it is related to social desirability bias. Hilliard et al. (2024) found that newer and more parameter LLMs exhibited a broader range of personality traits, including higher agreeableness, emotional stability, and openness. Salecha et al. (2024) also discovered that LLMs showed similar human-like social desirability biases when generating simulated data. To test this hypothesis, we conducted regression and correlation analyses between the BFI-2 and HEXACO-100 human responses, their LLM-generated responses, the differences between them, and the social desirability rating. The results are presented in Table 20 and Table 21 (for the regression line chart, see Appendix A, Figure 1 and Figure 2). Social desirability rating of the BFI-2 items were obtained from another ongoing study where 142 human resource practitioners were asked to rate how desirable each item was in general (1 = "Very undesirable", 2 = "Undesirable", 3 = "Slightly undesirable", 4 = "Neither desirable nor undesirable", 5 = "Slightly desirable", 6 = "Desirable", 7 = "Very desirable"). Another group of three PhD students and four PhDs with a psychology background rated the social desirability of each HEXACO-100 item using the same 7-point scale. Average ratings across all raters were used as the social desirability estimates of each item.

Both BFI-2 and HEXACO-100 results demonstrated a strong positive correlation between social desirability ratings and both LLM-generated responses and human responses, with correlation coefficients around .70 ± .10 for BFI-2 and .50 ± .20 for HEXACO-100. Regression analysis was conducted to explore the effect of social desirability ratings on these datasets. When the social desirability rating of an item was neutral (4), the predicted item mean for most LLM-generated responses was 3, indicating a neutral or no opinion stance. However, BFI-2 LLM-generated responses generated by the LLaMA3 model with the *persona* method showed a slightly lower predicted item mean, whereas HEXACO-100 LLM-generated responses generated by GPT models with the *shape* method showed a slightly higher predicted item mean.

And for the mean differences between human responses and LLM-generated responses, most exhibited positive correlations with social desirability ratings, except for the mean differences between human responses and data generated by the LLaMA3 *persona* method, which showed little to no correlation with social desirability ratings. When the social desirability rating was neutral (4), the mean difference for most data was around 0, except for the predicted mean difference with the *persona* LLaMA3 data, which was 0.40. In contrast, for HEXACO-100, the predicted mean differences by *shape* GPT3.5 and *shape* GPT4 were -0.33 and -0.19, respectively.

Line charts in Appendix A, Figure 1, and Figure 2, further illustrate that as the social desirability rating of an item increased, the item mean for both human responses and LLM-generated responses also increased. For most human responses and LLM-generated responses, the mean differences grew larger as the social desirability rating increased and smaller as it decreased. This indicates that when the social desirability rating was high, the item mean in human responses exceeded that in LLM-generated responses, and when the rating was low, the item mean in human responses was lower than that in LLM-generated responses. This suggests that human responses are more influenced by social desirability. However, the mean difference between human responses and data generated using the LLaMA3 *persona* method remained stable, with the lines in the chart appearing nearly parallel to the human responses line.

Table 20

Regression Analysis and Correlation of Social Desirability Ratings: Human Responses vs. BFI-2 LLM-Generated Responses

	human sample		*persona* GPT3.5		*shape* GPT3.5		*persona* GPT4		*shape* GPT4		*persona* LLaMA3		*shape* LLaMA3	
BFI-2	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept	1.15	0.23	1.56	0.18	1.99	0.14	2.45	0.10	2.33	0.09	0.71	0.21	2.25	0.11
Social desirability	0.46	0.05	0.37	0.04	0.26	0.03	0.12	0.02	0.17	0.02	0.47	0.05	0.14	0.02
R^2	.59		.60		.54		.35		.59		.65		.37	
Predicted score at neutral point	2.99		3.04		3.03		2.93		3.01		2.59		2.81	
Correlation	.77		.77		.73		.60		.77		.80		.61	
			$\text{MSD}_{\textit{persona}\ \text{GPT3.5}}$		$\text{MSD}_{\textit{shape}\ \text{GPT3.5}}$		$\text{MSD}_{\textit{persona}\ \text{GPT4}}$		$\text{MSD}_{\textit{shape}\ \text{GPT4}}$		$\text{MSD}_{\textit{persona}\ \text{LLaMA3}}$		$\text{MSD}_{\textit{shape}\ \text{LLaMA3}}$	
			B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept			-0.41	0.17	-0.85	0.21	-1.30	0.19	-1.19	0.18	0.44	0.21	-1.11	0.23
Social desirability			0.09	0.04	0.21	0.05	0.34	0.04	0.29	0.04	-0.01	0.05	0.33	0.05
R^2			.09		.25		.54		.47		.00		.42	
Predicted score at neutral point			-0.05		-0.01		0.06		-0.03		0.40		0.21	
Correlation			.30		.50		.74		.69		-.03		.65	

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. MSD = mean score difference.

Table 21

Regression Analysis and Correlation of Social Desirability Ratings: Human Responses vs. HEXACO-100 LLM-Generated Responses

	human sample		*persona* GPT3.5		*shape* GPT3.5		*persona* GPT4		*shape* GPT4		*persona* LLaMA3		*shape* LLaMA3	
HEXACO-100	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept	1.72	0.12	1.91	0.14	2.53	0.15	2.61	0.06	2.60	0.10	1.64	0.22	2.57	0.13
Social desirability	0.32	0.03	0.28	0.03	0.19	0.03	0.09	0.01	0.15	0.02	0.32	0.05	0.11	0.03
R^2	.55		.42		.24		.29		.28		.29		.13	
Predicted score at neutral point	3.00		3.03		3.29		2.97		3.20		2.92		3.01	
Correlation	.74		.65		.49		.54		.53		.54		.36	
			$\text{MSD}_{\textit{persona}\ \text{GPT3.5}}$		$\text{MSD}_{\textit{shape}\ \text{GPT3.5}}$		$\text{MSD}_{\textit{persona}\ \text{GPT4}}$		$\text{MSD}_{\textit{shape}\ \text{GPT4}}$		$\text{MSD}_{\textit{persona}\ \text{LLaMA3}}$		$\text{MSD}_{\textit{shape}\ \text{LLaMA3}}$	
			B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept			-0.19	0.13	-0.81	0.16	-0.88	0.11	-0.87	0.12	0.09	0.18	-0.85	0.16
Social desirability			0.04	0.03	0.12	0.04	0.22	0.03	0.17	0.03	-0.01	0.04	0.20	0.04
R^2			.02		.09		.43		.27		.00		.24	
Predicted score at neutral point			-0.03		-0.33		0.00		-0.19		0.05		-0.05	
Correlation			.13		.31		.65		.52		-.02		.49	

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. MSD = mean score difference.

Additional Analyses: Distinguishing Human vs. LLM Responses

In addition to differences in descriptive statistics and psychometric performance, the distinction between LLM-generated responses and human responses can be further examined by mixing the two types of data for a classification task. Specifically, we can use a logistic regression model to evaluate and identify potential differences between these two types of data.

Method

We used the Scikit-learn library (Pedregosa et al., 2011) to implement the logistic regression model on BFI-2 LLM-generated responses and human responses, as well as HEXACO-100 LLM-generated responses and human responses. LLM-generated responses were considered the positive class, while human responses were considered the negative class. To create a balanced analysis environment, we used equal representation by randomly selecting 300 samples each from human responses for both the BFI-2 and HEXACO-100 scales, considering that LLM-generated responses include only 300 samples for each. To avoid selection bias in the human samples, we randomly selected 300 samples from the entire pool 50 times, ensuring that all samples were included in the analysis. In another word, we employed a validation approach using five-fold cross-validation repeated over 50 iterations to enhance the reliability of our findings.

Each dataset comprised specific item-based features: 60 items from BFI-2 and 100 items from HEXACO-100. These items were used as input features for the logistic regression models. The regression models were evaluated separately for each LLM configuration to ascertain specific performance metrics. The outcomes of these analyses were quantitatively summarized in terms of mean and standard deviation for key performance metrics such as accuracy, precision, recall, and F1 score. These results have been comprehensively detailed in Table 22.

Results

For the BFI-2 data, there was a very clear pattern, the accuracy, precision, recall, and F1 score metrics of the *shape* method for the GPT series models were consistently lower than those of the *persona* method ($M_{\text{accuracy } persona \text{ GPT3.5}} = 0.90$ vs. $M_{\text{accuracy } shape \text{ GPT3.5}} = 0.74$, $M_{\text{precision } persona \text{ GPT3.5}} = 0.90$ vs. $M_{\text{precision } shape \text{ GPT3.5}} = 0.74$, $M_{\text{recall } persona \text{ GPT3.5}} = 0.90$ vs. $M_{\text{recall } shape \text{ GPT3.5}} = 0.74$, $M_{\text{F1 } persona \text{ GPT3.5}} = 0.90$ vs. $M_{\text{F1 } shape \text{ GPT3.5}} = 0.74$; $M_{\text{accuracy } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{accuracy } shape \text{ GPT4}} = 0.74$, $M_{\text{precision } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{precision } shape \text{ GPT4}} = 0.75$, $M_{\text{recall } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{recall } shape \text{ GPT4}} = 0.74$, $M_{\text{F1 } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{F1 } shape \text{ GPT4}} = 0.74$). This suggested that the *shape* method generates simulated data that is more similar to human responses than the *persona* method. However, the relatively high accuracy and other metrics still showed discernible differences compared to human responses. Additionally, the LLaMA3-generated LLM-generated responses for the BFI-2 personality scale were comparatively closer to human responses ($M_{\text{accuracy } persona \text{ LLaMA3}} = 0.65$ & $M_{\text{accuracy } shape \text{ LLaMA3}} = 0.64$, $M_{\text{precision } persona \text{ LLaMA3}} = 0.65$ & $M_{\text{precision } shape \text{ LLaMA3}} = 0.65$, $M_{\text{recall } persona \text{ LLaMA3}} = 0.65$ & $M_{\text{recall } shape \text{ LLaMA3}} = 0.64$, $M_{\text{F1 } persona \text{ LLaMA3}} = 0.65$ & $M_{\text{F1 } shape \text{ LLaMA3}} = 0.64$). The standard deviations of all results were very small, indicating the stability of these findings (see Table 20).

For HEXACO-100 data, a similar pattern existed where the accuracy and precision metrics of the *shape* method in the GPT series models were also lower than those of the *persona* method ($M_{\text{accuracy } persona \text{ GPT3.5}} = 0.86$ vs. $M_{\text{accuracy } shape \text{ GPT3.5}} = 0.62$, $M_{\text{precision } persona \text{ GPT3.5}} = 0.86$ vs. $M_{\text{precision } shape \text{ GPT3.5}} = 0.62$, $M_{\text{recall } persona \text{ GPT3.5}} = 0.86$ vs. $M_{\text{recall } shape \text{ GPT3.5}} = 0.62$, $M_{\text{F1 } persona \text{ GPT3.5}} = 0.86$ vs. $M_{\text{F1 } shape \text{ GPT3.5}} = 0.62$; $M_{\text{accuracy } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{accuracy } shape \text{ GPT4}} = 0.68$, $M_{\text{precision } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{precision } shape \text{ GPT4}} = 0.68$, $M_{\text{recall } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{recall } shape \text{ GPT4}} = 0.68$, $M_{\text{F1 } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{F1 } shape \text{ GPT4}} = 0.68$). However, the differences were larger compared to the BFI-2 data. This suggested that for the HEXACO-100 personality scale, the differences between

LLM-generated responses generated by the GPT series models using the *persona* and *shape* methods were larger, and HEXACO-100 GPT series models' *shape* methods data were closer to human responses. The relatively high accuracy and other metrics still showed discernible differences compared to human responses. Also, the LLaMA3-generated LLM-generated responses for the HEXACO-100 personality scale were comparatively closer to human responses ($M_{\text{accuracy } persona \text{ LLaMA3}} = 0.56$ & $M_{\text{accuracy } shape \text{ LLaMA3}} = 0.68$, $M_{\text{precision } persona \text{ LLaMA3}} = 0.56$ & $M_{\text{precision } shape \text{ LLaMA3}} = 0.68$, $M_{\text{recall } persona \text{ LLaMA3}} = 0.55$ & $M_{\text{recall } shape \text{ LLaMA3}} = 0.68$, $M_{\text{F1 } persona \text{ LLaMA3}} = 0.55$ & $M_{\text{F1 } shape \text{ LLaMA3}} = 0.68$). Additionally, the standard deviations of the results were very small, indicating that these results were stable (see Table 20).

In summary, the simulated data generated based on LLaMA3 were most similar to the human responses. The simulated data produced by the *shape* method were more similar to the human responses compared to the *persona* method. However, overall, there were still significant differences between the LLM-generated responses and the human responses.

Table 22

Classification Results between LLM-Generated Responses and Human Responses

LLM-Generated Responses	BFI-2								HEXACO-100							
	Accuracy		Precision		Recall		F1 score		Accuracy		Precision		Recall		F1 score	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
persona GPT3.5	0.90	0.02	0.90	0.02	0.90	0.02	0.90	0.02	0.86	0.03	0.86	0.03	0.86	0.03	0.86	0.03
shape GPT3.5	0.74	0.04	0.74	0.04	0.74	0.04	0.74	0.04	0.62	0.04	0.62	0.04	0.62	0.04	0.62	0.04
persona GPT4	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01
shape GPT4	0.74	0.04	0.75	0.03	0.74	0.04	0.74	0.04	0.68	0.03	0.68	0.03	0.68	0.03	0.68	0.03
persona LLaMA3	0.65	0.03	0.65	0.03	0.65	0.03	0.65	0.03	0.56	0.03	0.56	0.04	0.55	0.04	0.55	0.04
shape LLaMA3	0.64	0.03	0.65	0.04	0.64	0.03	0.64	0.04	0.68	0.03	0.68	0.03	0.68	0.03	0.68	0.03

Note. For BFI-2, n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses; for HEXACO-100, n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

General Discussion

This paper presents a comparative analysis of the descriptive statistics and psychometric performance of LLM-generated responses versus human responses on the BFI-2 and HEXACO-100. Furthermore, we investigated the relationship between social desirability ratings and both data types, as well as the classification accuracy of the two data types. These three experiments collectively revealed distinct differences between LLM-generated responses and human data and showed that LLMs cannot effectively simulate human data.

Our findings suggest that when responding to personality scales, LLM-generated responses differed distinctly from human responses. The standard deviation of LLM-generated responses was close to that of human responses in terms of *MAE*, but not in terms of profile correlation. This means that while the LLM-generated responses matched human data in overall variability, LLMs still fell short in capturing individual difference patterns. For psychometric performance, LLM-generated responses, whether it was the model of each domain or the overall Five-Factor model or Six-Factor model, showed poorer model fit and structural validity. Scale reliability at the facet level and domain level also exhibited obvious differences, and correlations with theoretically distinct factors in LLM-generated responses were more pronounced compared to human responses.

This difference was not reflected in the mean values of the personality higher-order structure, which was the greatest similarity between the two types of data. This also aligned with previous research indicating that LLMs can simulate human behavior (Huang, Wang, Lam, et al., 2023; Jiang et al., 2023). However, as previously stated, focusing solely on the means of the main domains of personality is insufficient; individuals respond to individual items, not the scales.

Exploring the social desirability rating further highlighted the differences between LLM-generated responses and human responses. Although there was a strong positive correlation, with a correlation coefficient of approximately 0.70 ± 0.10, between the scores on each item and the social desirability rating for both data types, the mean differences between LLM-generated responses and human responses increased as the social desirability rating rose and decreased as it fell. This indicates that when the social desirability rating is high, the mean item scores in the human responses are higher than those in the LLM-generated responses. Conversely, when the social desirability rating is low, the mean item scores in the human responses are lower than those in the LLM-generated responses. This suggests that human responses are more influenced and inflated by social desirability, whereas LLM-generated responses show constraints. Therefore, we believe that while the differences between LLM-generated responses and human responses are influenced by social desirability bias, the impact of this bias is greater on humans than on models. This results in more extreme responses from humans on items with high or low social desirability ratings, whereas the model's responses tend to be more neutral.

Finally, the results of the logistic regression analysis clearly showed differences between the LLM-generated responses and the human responses. By extracting the mean and standard deviation features at the item level, the regression model clearly demonstrated differences between the LLM-generated responses and the human responses. Across both BFI-2 and HEXACO-100 datasets, *persona* method data generally outperformed *shape* method data, with the LLaMA3 models showing the smallest differences, suggesting a closer approximation to human data. The analysis reveals persistent gaps between LLM-simulated and human responses, emphasizing the ongoing challenges in accurately modeling complex human traits with current AI technologies.

Models and Methods

Given the various differences between LLM-generated responses and human responses, our results have shown that simulated data generated by newer models with more parameters tend to be closer to human data. This observation may indicate the future full potential of LLM-generated responses. With continuous technological advancements and further model optimization, we have reason to believe that these differences will gradually diminish. Thus, we should maintain high expectations for the future capabilities of LLM-generated responses.

Moreover, although the two prompting methods discussed in this paper are widely used in the field of computer science, they seem to fall short in simulating human responses. The *persona* method attempts to simulate each individual but provides insufficient information. Considering the types of characters in literature, the concepts of flat characters and round characters might offer some insights for LLM simulation. Flat characters typically have one or two prominent traits, lacking depth and complexity, whereas round characters possess a rich inner world and multi-layered traits. If we provide more comprehensive simulated information to LLMs, such as ‘silicon persona’ (Argyle et al., 2023; Petrov et al., 2024), or refer to structured information input, such as a personality chatbot (Fan et al., 2023), the quality of LLM-generated responses might improve. The *shape* method increases data diversity to some extent, but its core principle is to view personality traits as uniformly distributed. However, in the real world, human responses’ personality traits usually follow a normal distribution (Kachur et al. 2020). This means that the method we use fails to adequately capture the natural variability of individual human personality traits. If we map using a normal distribution, the quality of LLM-generated responses might improve.

Limitations

One limitation of this study is the lack of exploration into how LLMs simulate data from different populations. We hypothesize that this might be a shortcoming of current LLM-generated responses. The outputs of LLMs are based on their training data and the models' structures. This raises several essential questions that are important to consider: Is the training data of LLMs sufficiently representative of the perspectives of various population identities? Could this adversely affect marginalized groups? As A. Wang et al. (2024) noted, if LLMs cannot adequately represent the perspectives of different population identities, they should not be considered capable substitutes for human participants. The risk of bias and misrepresentation is high if the training data lacks diversity, potentially leading to skewed or inaccurate personality simulations of minority groups. Next, we performed some fundamental psychometric analyses, but additional analyses are needed, particularly involving measures beyond what we have tested. Meanwhile, traditional psychometric tools may require moderation to align with the capabilities of LLMs so that the linguistic nuances are accurately captured and interpreted. Furthermore, we did not examine thoroughly the underlying reasons for observed differences between human data and LLM-generated responses. While we have attempted to explain some of these discrepancies with social desirability, future research should conduct a deeper exploration to fully understand the reasons behind.

References

Aher, G., Arriaga, R., I., & Kalai, A. T. (2022). Using large language models to simulate multiple humans and replicate human subject studies. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2208.10264

Ai, Y., He, Z., Zhang, Z., Zhu, W., Hao, H., Yu, K., Chen, L., & Wang, R. (2024). Is cognition and action consistent or not: investigating large language model's personality. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.14679

Anglim, J., Dunlop, P. D., Wee, S., Horwood, S., Wood, J. K., & Marty, A. (2022). Personality and intelligence: A meta-analysis. *Psychological Bulletin*, *148*(5–6), 301–336. https://doi.org/10.1037/bul0000373

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*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.14165

Caron, G., & Srivastava, S. (2022). Identifying and manipulating the personality traits of language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.10276

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, *108*(8), 1277–1299. https://doi.org/10.1037/apl0001082

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: a case study. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing System.* https://doi.org/10.1145/3544548.3580688

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 (Cornell University)*. https://doi.org/10.48550/arxiv.2402.08341

Huang, J., Wang, W., Lam, M. H., Li, E. J., Jiao, W., & Lyu, M. R. (2023). Revisiting the reliability of psychological scales on large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.19926

Huang, J., Wang, W., Li, E. J., Lam, M. H., Ren, S., Yuan, Y., Jiao, W., Tu, Z., & Lyu, M. R. (2023). Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.01386

Jiang, H., Zhang, X., Cao, X., Kabbara, J., & Roy, D. (2023). PersonaLLM: Investigating the ability of large language models to express big five personality traits. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.02547

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *Big Five Inventory (BFI)* [Database record]. APA PsycTests. https://doi.org/10.1037/t07550-000

Kachur, A., Osin, E., Davydov, D., Shutilov, K., & Novokshonov, A. (2020). Assessing the Big Five personality traits using real-life static facial images. Scientific Reports, 10(1), 8487.

Lampinen, A. K., Dasgupta, I., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*(7). https://doi.org/10.1093/pnasnexus/pgae233

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, *25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Lorenzo-Seva, U., & Berge, J. M. F. T. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57–64. https://doi.org/10.1027/1614-2241.2.2.57

Meta-Llama. (2024). *GitHub - meta-llama/llama3: The official Meta Llama 3 GitHub site*. GitHub. https://github.com/meta-llama/llama3

OpenAI. (2023). GPT-4 Technical Report. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.02155

Park, J., Min, B., Ma, X., & Kim, J. (2023). ChoiceMates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.01331

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231214460

Petrov, N. B., Serapio-García, G., & Rentfrow, J. (2024). Limited ability of LLMs to Simulate human psychological behaviours: A psychometric analysis. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.07248

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models show human-like social desirability biases in survey responses. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.06058

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.00184

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power.

Journal of Personality and Social Psychology, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLAMA: Open and efficient foundation language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13971

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). Large language models cannot replace human participants because they cannot portray identity groups. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.01908

Wang, P., Xiao, Z., Chen, H., & Oswald, F. L. (2024). Will the real Linda please stand up. . .To large language models? Examining the representativeness heuristic in LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.01461

Wang, X., Tu, Q., Fei, Y., Leng, Z., & Li, C. (2023). Does role-playing chatbots capture the character personalities? Assessing personality traits for role-playing chatbots. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.17976

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Liu, Q., Zhou, Y., Wang, W., Jiang, C., Zou, Y., . . . Gui, T. (2023). The rise and potential of large Language model based agents: A survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.07864

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1801.07243

Appendix A

BFI-2 Results

Table 23

Item Level Mean for BFI-2 Human Responses and LLM-Generated Responses

No.	Item Content	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
1	I am someone who Is outgoing, sociable	3.03	2.86	2.61	3.14	2.87	2.77	2.45
2	I am someone who Is compassionate, has a soft heart	4.34	4.12	3.53	3.19	3.22	3.87	3.17
3	I am someone who Tends to be disorganized	3.62	3.46	3.29	3.03	3.16	3.77	3.31
4	I am someone who Is relaxed, handles stress well	2.56	2.55	2.75	2.99	2.98	2.94	3.32
5	I am someone who Has few artistic interests	3.68	3.00	3.00	3.48	3.14	3.80	2.92
6	I am someone who Has an assertive personality	2.96	3.20	3.22	3.03	3.13	3.07	2.81
7	I am someone who Is respectful, treats others with respect	4.62	4.92	4.48	3.34	3.72	4.43	3.62
8	I am someone who Tends to be lazy	3.62	3.77	3.33	3.36	3.26	4.23	3.44
9	I am someone who Stays optimistic after experiencing a setback	2.44	2.12	2.56	2.87	2.83	2.29	2.92
10	I am someone who Is curious about many different things	4.43	4.21	3.79	3.36	3.25	3.71	3.15
11	I am someone who Rarely feels excited or eager	3.54	3.13	2.80	3.39	3.06	4.18	2.99
12	I am someone who Tends to find fault with others	3.34	3.53	3.43	3.03	3.24	4.31	3.46
13	I am someone who Is dependable, steady	4.33	3.83	3.20	3.25	3.40	3.70	3.11
14	I am someone who Is moody, has up and down mood swings	2.56	2.98	2.97	3.00	2.82	2.52	2.60
15	I am someone who Is inventive, finds clever ways to do things	3.90	3.52	3.32	3.18	3.26	3.24	2.91
16	I am someone who Tends to be quiet	2.35	2.97	2.88	3.01	2.95	3.44	3.25
17	I am someone who Feels little sympathy for others	3.71	3.28	3.45	3.67	3.44	4.47	3.54
18	I am someone who Is systematic, likes to keep things in order	3.96	3.42	3.12	3.12	3.29	2.46	2.59

19	I am someone who Can be tense	3.29	3.02	3.09	3.07	3.23	3.18	3.04
20	I am someone who Is fascinated by art, music, or literature	4.06	3.61	3.43	3.39	3.28	2.99	2.90
21	I am someone who Is dominant, acts as a leader	2.90	2.88	2.66	3.00	3.01	2.46	2.53
22	I am someone who Starts arguments with others	4.23	3.86	3.73	3.32	3.45	4.56	3.70
23	I am someone who Has difficulty getting started on tasks	3.31	3.08	2.83	3.06	2.99	3.85	3.23
24	I am someone who Feels secure, comfortable with self	2.26	2.52	2.53	2.88	2.55	2.24	2.65
25	I am someone who Avoids intellectual, philosophical discussions	3.93	3.10	2.57	3.14	3.16	3.36	3.13
26	I am someone who Is less active than other people	3.35	3.49	2.98	3.20	3.07	3.86	3.14
27	I am someone who Has a forgiving nature	3.78	3.27	3.70	3.01	3.29	3.48	3.28
28	I am someone who Can be somewhat careless	3.56	3.40	3.44	3.05	3.06	3.19	3.00
29	I am someone who Is emotionally stable, not easily upset	2.44	2.61	2.29	2.94	2.68	2.66	2.87
30	I am someone who Has little creativity	3.95	3.17	3.13	3.78	3.39	4.36	3.42
31	I am someone who Is sometimes shy, introverted	2.35	2.98	2.95	2.98	2.88	3.43	3.34
32	I am someone who Is helpful and unselfish with others	4.18	4.15	3.78	3.14	3.47	3.88	3.24
33	I am someone who Keeps things neat and tidy	3.72	3.45	3.33	3.00	3.43	2.47	2.63
34	I am someone who Worries a lot	3.30	2.63	2.87	3.03	3.01	2.59	2.50
35	I am someone who Values art and beauty	4.16	3.55	3.39	3.26	3.34	3.23	3.06
36	I am someone who Finds it hard to influence people	3.09	3.01	2.78	3.00	2.60	3.44	2.99
37	I am someone who Is sometimes rude to others	3.79	3.80	3.90	3.11	3.32	4.43	3.63
38	I am someone who Is efficient, gets things done	4.23	3.97	3.77	3.33	3.44	3.45	3.12
39	I am someone who Often feels sad	2.71	2.80	2.87	2.99	2.91	2.27	2.42
40	I am someone who Is complex, a deep thinker	4.02	3.84	3.36	3.24	3.58	3.59	3.14
41	I am someone who Is full of energy	3.14	3.25	2.96	3.18	3.15	3.39	2.86
42	I am someone who Is suspicious of others' intentions	2.75	3.08	3.23	3.03	2.93	3.91	3.40
43	I am someone who Is reliable, can always be counted on	4.32	4.05	3.59	3.24	3.50	3.76	3.18
44	I am someone who Keeps their emotions under control	2.22	2.66	2.47	2.99	2.48	2.92	2.90

45	I am someone who Has difficulty imagining things	4.17	3.06	3.11	3.71	3.34	4.08	3.43
46	I am someone who Is talkative	2.88	2.84	2.93	3.02	3.03	2.53	2.50
47	I am someone who Can be cold and uncaring	3.84	3.19	3.67	3.73	3.44	4.44	3.47
48	I am someone who Leaves a mess, doesn't clean up	4.15	4.30	3.71	3.53	3.46	4.25	3.25
49	I am someone who Rarely feels anxious or afraid	3.44	3.04	2.50	3.06	2.76	3.12	2.71
50	I am someone who Thinks poetry and plays are boring	3.66	3.65	3.15	3.04	3.03	3.85	3.15
51	I am someone who Prefers to have others take charge	2.96	3.52	3.32	3.19	3.24	4.33	3.20
52	I am someone who Is polite, courteous to others	4.50	4.57	3.86	3.15	3.44	4.13	3.56
53	I am someone who Is persistent, works until the task is finished	4.27	4.41	3.78	3.38	3.47	3.87	3.13
54	I am someone who Tends to feel depressed, blue	2.65	2.87	2.87	2.97	2.82	2.01	2.36
55	I am someone who Has little interest in abstract Ideas	3.83	2.94	2.70	3.06	3.03	3.40	2.93
56	I am someone who Shows a lot of Enthusiasm	3.35	3.71	2.98	3.39	3.08	4.08	2.81
57	I am someone who Assumes the best about people	3.37	3.43	3.43	3.02	3.21	3.40	3.13
58	I am someone who Sometimes behaves irresponsibly	3.64	3.41	3.72	2.97	2.94	3.44	3.22
59	I am someone who Is temperamental, gets emotional easily	2.53	2.72	2.77	2.98	2.63	2.33	2.66
60	I am someone who Is original, comes up with new Ideas	3.84	3.53	3.33	3.15	3.34	3.25	2.88

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Table 24

Item Level Standard Deviation for BFI-2 Human Responses and LLM-Generated Responses

No.	Item Content	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
1	I am someone who Is outgoing, sociable	1.40	0.91	1.14	0.65	1.08	1.31	1.37
2	I am someone who Is compassionate, has a soft heart	0.90	0.91	1.17	0.51	1.21	1.12	1.44
3	I am someone who Tends to be disorganized	1.36	0.60	1.11	0.27	1.01	0.95	1.46
4	I am someone who Is relaxed, handles stress well	1.30	0.83	1.16	0.39	1.21	1.16	1.39
5	I am someone who Has few artistic interests	1.33	0.58	0.84	0.68	0.95	1.31	1.32
6	I am someone who Has an assertive personality	1.38	0.64	1.03	0.27	1.02	1.18	1.37
7	I am someone who Is respectful, treats others with respect	0.61	0.41	1.04	0.65	1.29	0.92	1.42
8	I am someone who Tends to be lazy	1.31	0.70	1.17	0.74	1.29	1.04	1.50
9	I am someone who Stays optimistic after experiencing a setback	1.24	0.95	1.47	0.44	1.28	0.98	1.54
10	I am someone who Is curious about many different things	0.81	0.85	1.08	0.60	1.35	1.12	1.55
11	I am someone who Rarely feels excited or eager	1.26	0.40	1.10	0.70	1.15	0.75	1.41
12	I am someone who Tends to find fault with others	1.32	0.56	0.93	0.26	1.13	0.69	1.48
13	I am someone who Is dependable, steady	0.86	0.73	1.27	0.56	1.43	1.12	1.59
14	I am someone who Is moody, has up and down mood swings	1.36	0.37	1.05	0.26	1.16	1.09	1.41
15	I am someone who Is inventive, finds clever ways to do things	1.03	0.67	1.01	0.42	1.26	1.14	1.44
16	I am someone who Tends to be quiet	1.34	0.57	0.97	0.34	0.94	1.12	1.27
17	I am someone who Feels little sympathy for others	1.45	0.60	1.20	0.93	1.27	0.79	1.54
18	I am someone who Is systematic, likes to keep things in order	1.06	0.76	1.17	0.44	1.18	1.09	1.37
19	I am someone who Can be tense	1.27	0.46	0.99	0.34	0.90	1.08	1.20
20	I am someone who Is fascinated by art, music, or literature	1.17	0.94	1.31	0.66	1.07	1.44	1.36
21	I am someone who Is dominant, acts as a leader	1.35	0.63	1.01	0.40	1.14	1.15	1.49

22	I am someone who Starts arguments with others	1.05	0.84	1.24	0.68	1.26	0.77	1.54
23	I am someone who Has difficulty getting started on tasks	1.39	0.47	1.19	0.36	1.19	0.86	1.49
24	I am someone who Feels secure, comfortable with self	1.27	0.80	1.16	0.47	1.31	1.08	1.49
25	I am someone who Avoids intellectual, philosophical discussions	1.21	0.49	1.13	0.37	1.10	1.36	1.56
26	I am someone who Is less active than other people	1.31	0.79	1.12	0.56	1.01	1.04	1.31
27	I am someone who Has a forgiving nature	1.20	0.59	1.10	0.14	1.16	0.90	1.37
28	I am someone who Can be somewhat careless	1.25	0.55	0.85	0.36	1.19	1.11	1.38
29	I am someone who Is emotionally stable, not easily upset	1.27	0.77	1.08	0.43	1.13	1.01	1.47
30	I am someone who Has little creativity	1.20	0.56	0.89	0.93	1.19	0.69	1.32
31	I am someone who Is sometimes shy, introverted	1.36	0.64	0.89	0.42	1.04	1.11	1.30
32	I am someone who Is helpful and unselfish with others	0.88	0.76	1.13	0.46	1.20	1.05	1.50
33	I am someone who Keeps things neat and tidy	1.24	0.79	1.19	0.28	0.99	1.00	1.35
34	I am someone who Worries a lot	1.47	0.60	1.20	0.42	0.85	1.08	1.33
35	I am someone who Values art and beauty	1.07	0.86	1.02	0.56	1.01	1.30	1.35
36	I am someone who Finds it hard to influence people	1.20	0.19	0.84	0.23	1.10	1.04	1.37
37	I am someone who Is sometimes rude to others	1.25	0.98	1.15	0.45	1.27	0.91	1.49
38	I am someone who Is efficient, gets things done	0.88	0.83	1.25	0.62	1.23	1.12	1.46
39	I am someone who Often feels sad	1.43	0.55	0.83	0.37	1.00	0.98	1.25
40	I am someone who Is complex, a deep thinker	1.08	0.67	0.95	0.48	1.08	1.06	1.26
41	I am someone who Is full of energy	1.30	0.83	1.02	0.56	1.11	1.24	1.35
42	I am someone who Is suspicious of others’ intentions	1.30	0.40	1.10	0.23	1.10	0.83	1.49
43	I am someone who Is reliable, can always be counted on	0.88	0.84	1.31	0.53	1.39	1.09	1.59
44	I am someone who Keeps their emotions under control	1.14	0.63	0.95	0.22	0.97	1.08	1.44
45	I am someone who Has difficulty imagining things	1.06	0.28	0.83	0.81	1.10	0.53	1.27
46	I am someone who Is talkative	1.38	0.58	1.05	0.38	0.92	1.11	1.28
47	I am someone who Can be cold and uncaring	1.23	0.57	1.06	0.98	1.43	0.87	1.49

48	I am someone who Leaves a mess, doesn’t clean up	1.12	0.81	1.27	0.83	1.39	0.95	1.64
49	I am someone who Rarely feels anxious or afraid	1.37	0.50	1.16	0.31	0.91	1.12	1.37
50	I am someone who Thinks poetry and plays are boring	1.37	0.58	1.16	0.21	0.91	1.31	1.53
51	I am someone who Prefers to have others take charge	1.28	0.66	1.39	0.44	1.19	0.76	1.57
52	I am someone who Is polite, courteous to others	0.70	0.73	1.14	0.38	1.29	0.97	1.44
53	I am someone who Is persistent, works until the task is finished	0.91	0.74	1.16	0.65	1.30	1.13	1.48
54	I am someone who Tends to feel depressed, blue	1.46	0.54	0.99	0.38	1.14	1.07	1.37
55	I am someone who Has little interest in abstract Ideas	1.22	0.34	0.88	0.32	1.04	1.08	1.36
56	I am someone who Shows a lot of Enthusiasm	1.22	0.86	1.19	0.63	1.18	1.09	1.51
57	I am someone who Assumes the best about people	1.28	0.62	1.21	0.27	1.20	0.98	1.47
58	I am someone who Sometimes behaves irresponsibly	1.30	0.68	0.88	0.38	1.16	1.13	1.40
59	I am someone who Is temperamental, gets emotional easily	1.33	0.55	1.00	0.30	1.10	1.08	1.52
60	I am someone who Is original, comes up with new Ideas	1.07	0.69	1.08	0.40	1.21	1.13	1.41

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Table 25

Facet Level Mean for BFI-2 Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Sociability	2.65	2.91	2.84	3.04	2.93	3.04	2.89
Assertiveness	2.97	3.15	3.00	3.05	2.99	3.33	2.88
Energy Level	3.35	3.40	2.93	3.29	3.09	3.88	2.95
Compassion	4.02	3.69	3.61	3.43	3.39	4.16	3.36
Respectfulness	4.29	4.29	3.99	3.23	3.48	4.39	3.63
Trust	3.31	3.33	3.45	3.02	3.17	3.78	3.32
Organization	3.86	3.66	3.36	3.17	3.33	3.24	2.94
Productiveness	3.86	3.81	3.43	3.28	3.29	3.85	3.23
Responsibility	3.96	3.67	3.49	3.13	3.23	3.52	3.13
Anxiety	3.15	2.81	2.80	3.04	2.99	2.96	2.89
Depression	2.51	2.58	2.71	2.93	2.78	2.20	2.59
Emotional Volatility	2.44	2.74	2.62	2.98	2.65	2.61	2.76
Intellectual Curiosity	4.05	3.52	3.11	3.20	3.26	3.51	3.09
Aesthetic Sensitivity	3.89	3.45	3.24	3.29	3.20	3.47	3.01
Creative Imagination	3.96	3.32	3.22	3.46	3.33	3.73	3.16

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Table 26

Facet Level Standard Deviation for BFI-2 Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Sociability	1.17	0.51	0.84	0.39	0.87	0.98	1.00
Assertiveness	1.04	0.32	0.75	0.25	0.95	0.82	1.22
Energy Level	0.95	0.53	0.97	0.47	0.98	0.88	1.19
Compassion	0.81	0.45	0.92	0.59	1.15	0.81	1.36
Respectfulness	0.70	0.45	0.96	0.43	1.16	0.74	1.33
Trust	1.00	0.33	0.90	0.17	1.09	0.67	1.34
Organization	1.01	0.49	0.86	0.34	0.96	0.79	1.18
Productiveness	0.91	0.50	0.97	0.47	1.12	0.87	1.29
Responsibility	0.84	0.51	0.88	0.36	1.14	0.95	1.26
Anxiety	1.11	0.41	0.80	0.29	0.76	0.89	1.04
Depression	1.14	0.48	0.89	0.31	1.04	0.89	1.26
Emotional Volatility	1.10	0.40	0.76	0.23	0.94	0.92	1.25
Intellectual Curiosity	0.83	0.34	0.78	0.33	1.01	0.96	1.24
Aesthetic Sensitivity	1.01	0.49	0.88	0.45	0.88	1.09	1.26
Creative Imagination	0.88	0.35	0.73	0.49	1.06	0.74	1.17

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Table 27

Domain Level Mean for BFI-2 Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
E	2.99	3.15	2.92	3.13	3.00	3.42	2.90
A	3.87	3.77	3.68	3.23	3.35	4.11	3.43
C	3.89	3.71	3.43	3.19	3.28	3.54	3.10
N	2.70	2.71	2.71	2.98	2.81	2.59	2.75
O	3.97	3.43	3.19	3.32	3.26	3.57	3.09

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

Table 28

Domain Level Standard Deviation for BFI-2 Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
E	0.85	0.37	0.78	0.29	0.87	0.81	1.00
A	0.69	0.34	0.87	0.34	1.09	0.70	1.28
C	0.80	0.44	0.81	0.35	1.01	0.79	1.15
N	1.02	0.36	0.74	0.23	0.85	0.81	1.09
O	0.76	0.30	0.73	0.34	0.93	0.78	1.15

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

Table 29

Standardized Factor Loadings for BFI-2 Three-Factor Models of Each Domain

		human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sociability=~item1	.80	.67	.74	.75	.84	.72	.53
	Sociability=~item16	.82	.66	.84	.87	.79	.88	.75
	Sociability=~item31	.82	.75	.86	.83	.78	.87	.80
	Sociability=~item46	.76	.49	.57	.81	.92	.64	.57
	Assertiveness=~item6	.76	.73	.45	.71	.96	.84	.85
	Assertiveness=~item21	.89	.57	.40	.66	.87	.73	.69
	Assertiveness=~item36	.53	.11	.66	.69	.64	.72	.88
	Assertiveness=~item51	.69	.05	.54	.46	.75	.54	.73
	Energy Level=~item11	.45	.23	.87	.48	.80	.73	.89
	Energy Level=~item26	.54	.58	.82	.83	.82	.75	.78
	Energy Level=~item41	.83	.79	.81	.81	.88	.87	.76
E	Energy Level=~item56	.77	.67	.81	.58	.82	.78	.77
	Compassion=~item2	.73	.66	.78	.75	.93	.75	.74
	Compassion=~item17	.34	-.04	.55	.74	.83	.86	.97
	Compassion=~item32	.62	.86	.90	.70	.80	.74	.86
	Compassion=~item47	.75	.09	.65	.71	.90	.85	.98
	Respectfulness=~item7	.71	.64	.86	.86	.95	.80	.84
A	Respectfulness=~item22	.60	-.07	.66	.53	.67	.68	.87

	Respectfulness=~item37	.70	.20	.73	.54	.86	.79	.94
	Respectfulness=~item52	.70	.83	.87	.84	.98	.80	.83
	Trust=~item12	.69	-.01	.67	.93	.86	.57	.94
	Trust=~item27	.68	-.59	.83	.47	.99	.80	.84
	Trust=~item42	.67	-.26	.71	.68	.91	.63	.95
	Trust=~item57	.78	-.69	.85	.37	.98	.82	.85
	Organization=~item3	.86	.41	.55	.61	.67	.75	.48
	Organization=~item18	.70	.82	.68	.81	.86	.66	.90
	Organization=~item33	.86	.84	.83	.61	.83	.68	.95
	Organization=~item48	.74	.09	.46	.50	.77	.74	.55
	Productiveness=~item8	.72	.46	.56	.60	.76	.68	.67
	Productiveness=~item23	.75	.42	.58	.38	.81	.65	.70
	Productiveness=~item38	.73	.79	.86	.82	.89	.88	.82
	Productiveness=~item53	.70	.64	.90	.86	.96	.82	.96
	Responsibility=~item13	.68	.76	.84	.90	.97	.94	.94
	Responsibility=~item28	.73	.34	.52	.45	.65	.58	.46
	Responsibility=~item43	.70	.83	.90	.90	.93	.95	.97
C	Responsibility=~item58	.65	.42	.64	.46	.73	.66	.63
	Anxiety=~item4	.85	.68	.52	.67	.79	.72	.62
	Anxiety=~item19	.69	.53	.70	.79	.80	.61	.72
	Anxiety=~item34	.79	.38	.60	.75	.85	.81	.79
	Anxiety=~item49	.66	.47	.55	.70	.42	.74	.70
N	Depression=~item9	.64	.61	.75	.41	.72	.68	.77

	Depression=~item24	.67	.74	.77	.46	.82	.86	.86
	Depression=~item39	.90	.40	.64	.86	.88	.81	.86
	Depression=~item54	.91	.27	.66	.93	.94	.88	.95
	Emotional Volatility=~item14	.80	.39	.75	1.06	.86	.86	.88
	Emotional Volatility=~item29	.88	.72	.60	.33	.81	.86	.85
	Emotional Volatility=~item44	.77	.60	.45	.34	.79	.73	.59
	Emotional Volatility=~item59	.80	.35	.67	.74	.77	.80	.87
	Intellectual Curiosity=~item10	.60	.64	.78	.78	.80	.81	.91
	Intellectual Curiosity=~item25	.63	-.05	.73	.56	.89	.76	.89
	Intellectual Curiosity=~item40	.65	.48	.55	.64	.83	.70	.57
	Intellectual Curiosity=~item55	.74	.08	.67	.47	.86	.82	.89
	Aesthetic Sensitivity=~item5	.66	.20	.60	.89	.78	.85	.91
	Aesthetic Sensitivity=~item20	.86	.81	.85	.77	.91	.85	.83
	Aesthetic Sensitivity=~item35	.87	.86	.84	.85	.92	.88	.81
	Aesthetic Sensitivity=~item50	.65	.03	.68	.38	.83	.41	.91
	Creative Imagination=~item15	.72	.64	.85	.61	.80	.89	.85
	Creative Imagination=~item30	.77	.14	.43	.69	.90	.63	.81
	Creative Imagination=~item45	.66	.03	.59	.57	.91	.49	.77
O	Creative Imagination=~item60	.80	.73	.81	.68	.79	.95	.81

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness.

Table 30

Standardized Factor Loadings for BFI-2 Five-Factor Model

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
E=~Sociability	.66	.59	.88	.48	.90	.85	.87
E=~Assertiveness	.59	.63	.80	.47	.88	.86	.73
E=~Energy Level	.77	.80	.93	.86	.94	.86	.85
A=~Compassion	.70	.73	.87	.76	.93	.90	.99
A=~Respectfulness	.80	.65	.83	.90	.92	.94	.85
A=~Trust	.65	.74	.98	.51	.96	.87	.95
C=~Organization	.70	.74	.82	.80	.93	.87	.97
C=~Productiveness	.89	.78	.59	.82	.86	.75	.81
C=~Responsibility	.79	.90	1.03	.89	.94	.95	.92
N=~Anxiety	.84	.71	.82	.67	.82	.75	.86
N=~Depression	.88	.92	.99	.88	1.03	.96	.96
N=~Emotional Volatility	.85	.56	.66	.64	.74	.79	.80
O=~Intellectual Curiosity	.74	.63	.90	.55	.97	.66	.93
O=~Aesthetic Sensitivity	.71	.60	.94	.50	.87	.65	.96
O=~Creative Imagination	.81	.65	.78	1.05	.92	.89	.82

Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

HEXACO Results

Table 31

Item Level Mean for HEXACO Human Responses and LLM-Generated Responses

No.	Item content	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
1	I would be quite bored by a visit to an art gallery.	3.63	3.31	2.54	3.10	3.03	3.58	2.85
2	I clean my office or home quite frequently.	3.39	2.92	3.07	2.99	3.32	3.02	2.97
3	I rarely hold a grudge, even against people who have badly wronged me.	3.06	3.21	3.81	3.03	3.47	3.78	3.53
4	I feel reasonably satisfied with myself overall.	3.70	3.36	3.46	3.28	3.54	3.92	3.45
5	I would feel afraid if I had to travel in bad weather conditions.	2.48	3.05	3.36	3.00	3.32	2.95	2.79
6	If I want something from a person I dislike, I will act very nicely toward that person in order to get it.	3.09	3.14	2.56	3.00	3.04	3.54	3.04
7	I am interested in learning about the history and politics of other countries.	3.44	3.17	3.66	3.02	3.27	2.55	3.04
8	When working, I often set ambitious goals for myself.	3.92	3.64	3.46	3.12	3.35	3.55	3.04
9	People sometimes tell me that I am too critical of others.	3.34	3.03	3.17	3.01	2.73	3.63	3.41
10	I rarely express my opinions in group meetings.	3.55	3.05	2.62	3.00	2.75	3.88	3.20
11	I sometimes can not help worrying about little things.	3.57	3.03	3.11	3.09	3.38	3.72	2.93
12	If I knew that I could never get caught, I would be willing to steal a million dollars.	3.87	4.43	4.27	3.48	3.62	4.73	3.68
13	I would like a job that requires following a routine rather than being creative.	3.61	3.58	2.76	3.36	3.13	3.86	3.18
14	I often check my work over repeatedly to find any mistakes.	3.46	3.35	3.31	3.05	3.28	3.31	3.00
15	People sometimes tell me that I'm too stubborn.	2.99	2.91	3.07	2.97	2.45	2.35	3.14
16	I avoid making “small talk” with people.	3.25	3.02	2.49	2.97	2.59	3.37	2.73
17	When I suffer from a painful experience, I need someone to make me feel comfortable.	3.08	3.49	3.61	3.03	3.24	3.86	3.24

18	Having a lot of money is not especially important to me.	3.13	3.27	4.15	3.03	3.63	3.94	3.27
19	I think that paying attention to radical ideas is a waste of time.	3.62	3.08	2.78	3.03	2.82	3.86	2.98
20	I make decisions based on the feeling of the moment rather than on careful thought.	3.55	3.41	2.72	2.95	2.96	3.04	2.85
21	People think of me as someone who has a quick temper.	3.74	3.09	3.18	3.06	3.34	4.31	3.80
22	I am energetic nearly all the time.	3.18	2.99	2.75	3.10	3.00	3.29	2.62
23	I feel like crying when I see other people crying.	3.07	3.24	3.52	3.00	3.13	2.89	2.78
24	I am an ordinary person who is no better than others.	3.27	3.00	3.09	2.98	3.42	3.70	3.53
25	I wouldn't spend my time reading a book of poetry.	3.20	3.16	2.41	2.95	2.51	2.64	2.51
26	I plan ahead and organize things, to avoid scrambling at the last minute.	3.73	4.37	3.41	3.12	3.34	2.99	2.68
27	My attitude toward people who have treated me badly is “forgive and forget”.	3.15	2.95	3.27	2.99	3.13	3.02	2.93
28	I think that most people like some aspects of my personality.	4.07	3.20	3.22	3.13	3.40	4.17	3.32
29	I don’t mind doing jobs that involve dangerous work.	2.90	3.41	3.31	2.95	3.01	3.55	3.31
30	I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed.	3.53	3.54	4.46	3.04	3.98	4.32	3.80
31	I enjoy looking at maps of different places.	3.54	3.08	3.24	3.02	3.48	3.13	3.08
32	I often push myself very hard when trying to achieve a goal.	3.99	3.82	3.51	3.24	3.48	3.81	2.93
33	I generally accept people’s faults without complaining about them.	3.37	4.12	3.90	3.04	3.24	3.61	3.26
34	In social situations, I'm usually the one who makes the first move.	3.02	2.95	2.68	2.96	2.92	2.97	2.78
35	I worry a lot less than most people do.	3.36	2.87	1.77	2.98	2.46	2.49	2.50
36	I would be tempted to buy stolen property if I were financially tight.	4.11	4.52	3.99	4.08	3.84	4.89	3.92
37	I would enjoy creating a work of art, such as a novel, a song, or a painting.	3.74	3.69	3.65	3.21	3.20	3.17	2.97
38	When working on something, I don't pay much attention to small details.	3.84	3.15	2.49	3.12	3.07	3.91	2.87
39	I am usually quite flexible in my opinions when people disagree with me.	3.44	3.27	3.24	2.99	3.00	2.92	2.85
40	I enjoy having lots of people around to talk with.	3.37	2.91	2.82	2.99	2.89	2.44	2.78
41	I can handle difficult situations without needing emotional support from anyone else.	2.73	2.52	2.20	2.99	2.70	2.25	2.71

42	I would like to live in a very expensive, high-class neighborhood.	3.14	3.53	3.16	3.04	2.81	3.93	3.27
43	I like people who have unconventional views.	3.80	3.44	3.46	3.09	3.29	4.19	3.41
44	I make a lot of mistakes because I don't think before I act.	3.81	3.31	3.02	3.08	3.14	3.65	3.03
45	I rarely feel anger, even when people treat me quite badly.	2.64	3.00	3.24	2.98	2.97	2.99	3.14
46	On most days, I feel cheerful and optimistic.	3.78	3.49	3.14	3.06	3.05	3.85	3.09
47	When someone I know well is unhappy, I can almost feel that person's pain myself.	3.68	3.29	3.48	3.21	3.42	4.13	3.37
48	I would not want people to treat me as though I were superior to them.	3.82	3.59	4.64	3.34	4.09	4.40	3.93
49	If I had the opportunity, I would like to attend a classical music concert.	3.52	3.01	3.20	3.00	3.22	2.35	2.85
50	People often joke with me about the messiness of my room or desk.	3.62	3.02	2.98	3.01	3.11	3.17	2.77
51	If someone has cheated me once, I will always feel suspicious of that person.	2.29	2.90	2.55	2.73	2.00	2.40	2.76
52	I feel that I am an unpopular person.	3.56	3.02	3.03	3.08	2.97	4.09	3.50
53	When it comes to physical danger, I am very fearful.	2.92	2.67	3.08	2.94	2.95	1.97	2.49
54	If I want something from someone, I will laugh at that person's worst jokes.	3.81	3.23	2.99	3.04	3.44	2.82	2.65
55	I would be very bored by a book about the history of science and technology.	3.43	3.30	2.45	3.01	2.91	3.18	2.84
56	Often when I set a goal, I end up quitting without having reached it.	3.72	3.49	2.67	3.03	3.01	3.83	3.13
57	I tend to be lenient in judging other people.	3.38	3.23	3.18	3.00	3.18	3.32	3.08
58	When I am in a group of people, I'm often the one who speaks on behalf of the group.	3.02	2.88	3.27	2.96	3.14	2.59	2.63
59	I rarely, if ever, have trouble sleeping due to stress or anxiety.	3.06	2.96	2.19	3.05	2.79	3.66	2.77
60	I would never accept a bribe, even if it were very large.	3.87	4.91	4.79	3.43	4.16	4.29	3.82
61	People have often told me that I have a good imagination.	3.72	3.38	3.31	3.13	3.12	3.45	3.14
62	I always try to be accurate in my work, even at the expense of time.	3.75	4.10	3.82	3.33	3.73	3.97	3.46
63	When people tell me that I’m wrong, my first reaction is to argue with them.	3.18	3.62	2.84	2.99	2.82	3.47	3.14
64	I prefer jobs that involve active social interaction to those that involve working alone.	3.52	2.99	3.18	3.07	2.89	3.18	2.99
65	Whenever I feel worried about something, I want to share my concern with another person.	3.52	3.17	3.42	3.01	3.08	3.13	3.01

66	I would like to be seen driving around in a very expensive car.	3.52	3.70	3.11	3.06	2.83	3.90	3.24
67	I think of myself as a somewhat eccentric person.	2.91	3.55	3.36	3.27	3.03	3.81	3.03
68	I don’t allow my impulses to govern my behavior.	3.29	3.59	4.01	3.06	3.92	2.88	3.18
69	Most people tend to get angry more quickly than I do.	3.48	3.03	3.79	3.07	3.77	3.97	3.93
70	People often tell me that I should try to cheer up.	3.72	3.08	2.89	3.02	3.05	3.31	3.08
71	I feel strong emotions when someone close to me is going away for a long time.	3.58	3.30	3.75	3.07	3.36	4.03	3.28
72	I think that I am entitled to more respect than the average person is.	3.75	3.17	2.88	3.00	3.08	3.73	3.47
73	Sometimes I like to just watch the wind as it blows through the trees.	3.78	3.24	3.66	3.10	3.50	2.95	3.07
74	When working, I sometimes have difficulties due to being disorganized.	3.53	3.10	2.79	3.02	3.00	3.25	3.12
75	I find it hard to fully forgive someone who has done something mean to me.	2.83	3.08	2.62	3.00	2.60	2.40	2.70
76	I sometimes feel that I am a worthless person.	3.57	3.34	3.25	3.52	3.52	4.38	3.71
77	Even in an emergency I wouldn't feel like panicking.	2.92	2.77	1.49	2.96	2.23	2.71	2.42
78	I would not pretend to like someone just to get that person to do favors for me.	3.65	4.31	4.50	3.28	4.16	4.78	4.08
79	I’ve never really enjoyed looking through an encyclopedia.	3.63	3.22	2.67	3.01	2.77	2.03	1.99
80	I do only the minimum amount of work needed to get by.	3.95	3.99	2.71	3.50	3.04	4.05	3.15
81	Even when people make a lot of mistakes, I rarely say anything negative.	2.95	3.23	3.69	3.05	3.40	3.50	3.28
82	I tend to feel quite self-conscious when speaking in front of a group of people.	2.70	3.31	2.75	2.99	2.82	3.69	3.36
83	I get very anxious when waiting to hear about an important decision.	3.73	3.10	3.27	3.02	3.23	2.76	2.46
84	I’d be tempted to use counterfeit money, if I were sure I could get away with it.	4.08	4.92	4.19	4.11	3.68	4.81	3.84
85	I don't think of myself as the artistic or creative type.	3.32	3.00	2.85	3.25	2.94	2.95	2.36
86	People often call me a perfectionist.	3.20	3.05	3.01	3.01	3.14	2.74	2.47
87	I find it hard to compromise with people when I really think I’m right.	2.81	2.69	2.22	2.98	2.55	2.50	2.52
88	The first thing that I always do in a new place is to make friends.	3.20	3.07	3.04	2.94	2.75	1.83	2.30
89	I rarely discuss my problems with other people.	3.20	3.01	2.48	2.96	2.52	3.19	2.57
90	I would get a lot of pleasure from owning expensive luxury goods.	3.34	3.61	3.36	3.02	2.80	4.03	3.28
91	I find it boring to discuss philosophy.	3.82	3.43	2.61	3.01	2.85	3.59	2.98

92	I prefer to do whatever comes to mind, rather than stick to a plan.	3.29	3.17	2.54	2.99	2.83	2.92	2.73
93	I find it hard to keep my temper when people insult me.	3.18	3.03	2.99	2.98	2.86	3.56	3.22
94	Most people are more upbeat and dynamic than I generally am.	3.36	3.29	3.13	2.95	2.83	2.56	2.30
95	I remain unemotional even in situations where most people get very sentimental.	3.50	3.11	2.89	3.06	2.91	3.68	2.91
96	I want people to know that I am an important person of high status.	3.69	3.11	3.32	3.05	2.92	3.89	3.43
97	I have sympathy for people who are less fortunate than I am.	4.12	4.64	4.30	3.16	3.73	4.21	3.56
98	I try to give generously to those in need.	3.66	3.62	3.88	3.09	3.47	3.47	3.13
99	It would not bother me to harm someone I didn't like.	4.08	4.72	4.14	4.04	4.05	4.87	4.10
100	People see me as a hard-hearted person.	3.77	3.05	3.35	3.13	3.20	4.43	3.66

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Table 32

Item Level Standard Deviation for HEXACO Human Responses and LLM-Generated Responses

No.	Item content	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
1	I would be quite bored by a visit to an art gallery.	1.10	0.54	1.26	0.34	1.02	0.98	1.44
2	I clean my office or home quite frequently.	1.09	0.51	1.16	0.24	1.03	1.06	1.39
3	I rarely hold a grudge, even against people who have badly wronged me.	1.12	0.52	1.11	0.24	1.18	0.87	1.44
4	I feel reasonably satisfied with myself overall.	0.92	0.64	1.15	0.55	1.28	0.90	1.40
5	I would feel afraid if I had to travel in bad weather conditions.	1.16	0.30	1.05	0.19	0.92	1.09	1.39
6	If I want something from a person I dislike, I will act very nicely toward that person in order to get it.	1.07	0.54	1.01	0.13	1.14	1.17	1.56
7	I am interested in learning about the history and politics of other countries.	1.13	0.63	1.07	0.14	1.07	0.99	1.48
8	When working, I often set ambitious goals for myself.	0.88	0.63	1.21	0.44	1.22	1.21	1.47
9	People sometimes tell me that I am too critical of others.	1.08	0.28	0.93	0.18	0.95	0.77	1.12
10	I rarely express my opinions in group meetings.	1.06	0.31	1.21	0.30	1.17	0.91	1.54
11	I sometimes can not help worrying about little things.	1.10	0.42	1.22	0.37	1.13	0.89	1.37
12	If I knew that I could never get caught, I would be willing to steal a million dollars.	1.19	0.91	1.27	0.84	1.34	0.83	1.62
13	I would like a job that requires following a routine rather than being creative.	1.06	0.69	1.29	0.56	1.05	1.09	1.44
14	I often check my work over repeatedly to find any mistakes.	1.04	0.58	1.23	0.42	1.15	1.25	1.51
15	People sometimes tell me that I'm too stubborn.	1.14	0.42	0.97	0.17	0.84	0.85	1.11
16	I avoid making “small talk” with people.	1.11	0.45	1.18	0.28	0.98	1.09	1.41
17	When I suffer from a painful experience, I need someone to make me feel comfortable.	1.13	0.75	1.09	0.24	1.11	0.97	1.31
18	Having a lot of money is not especially important to me.	1.07	0.60	1.01	0.35	0.93	0.96	1.38
19	I think that paying attention to radical ideas is a waste of time.	0.92	0.43	1.22	0.17	0.95	0.59	1.41

20	I make decisions based on the feeling of the moment rather than on careful thought.	0.97	0.67	1.36	0.30	1.40	1.22	1.63
21	People think of me as someone who has a quick temper.	1.08	0.31	0.82	0.35	1.27	0.97	1.37
22	I am energetic nearly all the time.	1.07	0.68	1.08	0.49	1.06	1.36	1.38
23	I feel like crying when I see other people crying.	1.15	0.64	1.26	0.17	1.11	1.15	1.43
24	I am an ordinary person who is no better than others.	1.12	0.18	0.98	0.33	1.14	0.98	1.37
25	I wouldn't spend my time reading a book of poetry.	1.27	0.44	1.13	0.28	0.95	1.27	1.46
26	I plan ahead and organize things, to avoid scrambling at the last minute.	1.04	0.88	1.58	0.55	1.29	1.29	1.45
27	My attitude toward people who have treated me badly is "forgive and forget".	1.04	0.37	1.19	0.23	1.16	1.16	1.47
28	I think that most people like some aspects of my personality.	0.64	0.51	1.06	0.35	1.22	0.70	1.31
29	I don't mind doing jobs that involve dangerous work.	1.14	0.91	1.14	0.39	1.01	1.38	1.55
30	I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed.	1.15	0.91	0.98	0.22	0.88	0.93	1.40
31	I enjoy looking at maps of different places.	1.12	0.34	0.90	0.20	0.69	1.06	1.25
32	I often push myself very hard when trying to achieve a goal.	0.88	0.89	1.30	0.57	1.23	1.19	1.48
33	I generally accept people's faults without complaining about them.	0.94	0.80	1.14	0.27	1.19	0.88	1.36
34	In social situations, I'm usually the one who makes the first move.	1.04	0.42	1.09	0.41	1.13	1.30	1.48
35	I worry a lot less than most people do.	1.13	0.53	0.97	0.37	1.14	1.09	1.35
36	I would be tempted to buy stolen property if I were financially tight.	0.99	0.88	1.41	1.00	1.31	0.47	1.50
37	I would enjoy creating a work of art, such as a novel, a song, or a painting.	1.17	0.84	1.33	0.49	1.14	1.33	1.56
38	When working on something, I don't pay much attention to small details.	0.96	0.86	1.15	0.43	1.28	0.87	1.48
39	I am usually quite flexible in my opinions when people disagree with me.	0.93	0.49	0.93	0.14	1.18	1.08	1.50
40	I enjoy having lots of people around to talk with.	1.07	0.71	1.26	0.46	1.11	1.05	1.48
41	I can handle difficult situations without needing emotional support from anyone else.	1.08	0.68	1.18	0.23	1.11	0.97	1.45
42	I would like to live in a very expensive, high-class neighborhood.	1.12	0.76	1.26	0.29	0.96	1.16	1.43
43	I like people who have unconventional views.	0.77	0.67	1.16	0.31	1.07	0.78	1.36

44	I make a lot of mistakes because I don't think before I act.	0.93	0.57	1.08	0.40	1.39	0.95	1.51
45	I rarely feel anger, even when people treat me quite badly.	1.09	0.19	0.86	0.13	0.97	1.11	1.42
46	On most days, I feel cheerful and optimistic.	0.94	0.83	1.36	0.50	1.28	1.08	1.55
47	When someone I know well is unhappy, I can almost feel that person's pain myself.	0.93	0.51	1.18	0.48	1.26	0.96	1.40
48	I would not want people to treat me as though I were superior to them.	0.92	0.94	0.81	0.62	0.82	0.83	1.22
49	If I had the opportunity, I would like to attend a classical music concert.	1.20	0.55	1.19	0.26	0.96	1.03	1.27
50	People often joke with me about the messiness of my room or desk.	1.19	0.23	1.22	0.17	0.99	1.20	1.40
51	If someone has cheated me once, I will always feel suspicious of that person.	0.97	0.41	0.95	0.47	0.90	0.88	1.31
52	I feel that I am an unpopular person.	1.04	0.21	0.89	0.41	1.13	0.83	1.33
53	When it comes to physical danger, I am very fearful.	1.13	0.61	1.34	0.34	0.97	0.70	1.39
54	If I want something from someone, I will laugh at that person's worst jokes.	0.94	0.50	1.04	0.21	1.29	1.19	1.52
55	I would be very bored by a book about the history of science and technology.	1.19	0.59	1.16	0.24	0.91	1.17	1.46
56	Often when I set a goal, I end up quitting without having reached it.	0.99	0.53	1.29	0.23	1.30	0.86	1.50
57	I tend to be lenient in judging other people.	0.94	0.55	1.00	0.13	1.21	0.95	1.42
58	When I am in a group of people, I'm often the one who speaks on behalf of the group.	1.04	0.55	1.15	0.34	1.19	1.26	1.44
59	I rarely, if ever, have trouble sleeping due to stress or anxiety.	1.28	0.33	1.14	0.24	0.93	0.99	1.35
60	I would never accept a bribe, even if it were very large.	1.12	0.45	0.80	0.77	0.91	0.99	1.37
61	People have often told me that I have a good imagination.	0.94	0.62	0.87	0.42	1.21	1.12	1.31
62	I always try to be accurate in my work, even at the expense of time.	0.90	0.79	0.99	0.59	1.12	1.06	1.39
63	When people tell me that I'm wrong, my first reaction is to argue with them.	1.02	0.60	1.09	0.21	1.26	1.05	1.64
64	I prefer jobs that involve active social interaction to those that involve working alone.	1.09	0.80	1.36	0.62	1.21	1.51	1.51
65	Whenever I feel worried about something, I want to share my concern with another person.	1.01	0.49	1.29	0.26	1.21	1.21	1.48
66	I would like to be seen driving around in a very expensive car.	1.16	0.86	1.30	0.37	1.08	1.18	1.60
67	I think of myself as a somewhat eccentric person.	1.09	0.64	0.91	0.53	0.90	1.18	1.35

68	I don’t allow my impulses to govern my behavior.	0.99	0.76	0.96	0.42	1.07	1.15	1.43
69	Most people tend to get angry more quickly than I do.	0.95	0.31	0.82	0.27	1.06	0.59	1.02
70	People often tell me that I should try to cheer up.	1.02	0.55	1.09	0.27	1.16	1.26	1.50
71	I feel strong emotions when someone close to me is going away for a long time.	1.04	0.58	1.11	0.32	1.16	0.90	1.37
72	I think that I am entitled to more respect than the average person is.	1.01	0.45	1.23	0.15	1.07	1.13	1.62
73	Sometimes I like to just watch the wind as it blows through the trees.	0.98	0.59	1.24	0.36	0.81	1.28	1.58
74	When working, I sometimes have difficulties due to being disorganized.	1.11	0.49	1.15	0.29	1.07	1.05	1.41
75	I find it hard to fully forgive someone who has done something mean to me.	1.10	0.46	0.90	0.20	1.01	0.94	1.36
76	I sometimes feel that I am a worthless person.	1.23	0.72	1.26	0.90	1.37	0.94	1.38
77	Even in an emergency I wouldn't feel like panicking.	1.13	0.62	0.81	0.25	1.07	1.15	1.33
78	I would not pretend to like someone just to get that person to do favors for me.	1.09	1.08	1.19	0.61	0.89	0.69	1.25
79	I’ve never really enjoyed looking through an encyclopedia.	1.07	0.53	1.06	0.08	0.79	0.74	0.87
80	I do only the minimum amount of work needed to get by.	0.95	1.06	1.61	0.86	1.54	1.10	1.68
81	Even when people make a lot of mistakes, I rarely say anything negative.	0.99	0.54	1.16	0.25	1.29	0.99	1.49
82	I tend to feel quite self-conscious when speaking in front of a group of people.	1.19	0.54	1.13	0.30	1.13	0.96	1.43
83	I get very anxious when waiting to hear about an important decision.	0.97	0.38	1.16	0.26	0.91	1.09	1.34
84	I’d be tempted to use counterfeit money, if I were sure I could get away with it.	1.08	0.39	1.46	0.98	1.43	0.71	1.54
85	I don't think of myself as the artistic or creative type.	1.16	0.50	0.78	0.54	0.91	1.38	1.16
86	People often call me a perfectionist.	1.11	0.37	0.83	0.27	1.01	1.11	1.22
87	I find it hard to compromise with people when I really think I’m right.	1.03	0.79	0.97	0.16	0.97	1.04	1.34
88	The first thing that I always do in a new place is to make friends.	1.02	0.43	1.31	0.43	1.25	0.71	1.44
89	I rarely discuss my problems with other people.	1.12	0.35	1.08	0.26	0.98	1.17	1.44
90	I would get a lot of pleasure from owning expensive luxury goods.	1.17	0.72	1.24	0.31	0.99	1.09	1.53
91	I find it boring to discuss philosophy.	1.06	0.54	1.14	0.10	0.83	0.96	1.34
92	I prefer to do whatever comes to mind, rather than stick to a plan.	1.01	0.74	1.35	0.35	1.40	1.24	1.60
93	I find it hard to keep my temper when people insult me.	1.11	0.48	1.01	0.16	1.14	0.98	1.54

94	Most people are more upbeat and dynamic than I generally am.	1.02	0.47	1.03	0.40	1.22	1.08	1.20
95	I remain unemotional even in situations where most people get very sentimental.	1.08	0.49	1.18	0.26	1.08	0.90	1.38
96	I want people to know that I am an important person of high status.	1.05	0.51	1.24	0.38	1.14	1.32	1.63
97	I have sympathy for people who are less fortunate than I am.	0.75	0.73	1.22	0.51	1.21	0.85	1.36
98	I try to give generously to those in need.	0.85	0.79	1.29	0.33	1.18	1.03	1.48
99	It would not bother me to harm someone I didn't like.	1.01	0.76	1.50	1.00	1.27	0.63	1.49
100	People see me as a hard-hearted person.	1.00	0.36	1.23	0.48	1.22	0.78	1.40

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Table 33

Facet Level Mean for HEXACO Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Sincerity	3.52	3.56	3.63	3.09	3.66	3.86	3.39
Fairness	3.98	4.69	4.31	3.77	3.82	4.68	3.82
Greed Avoidance	3.28	3.53	3.44	3.04	3.02	3.95	3.27
Modesty	3.64	3.22	3.48	3.09	3.38	3.93	3.59
Fearfulness	2.81	2.97	2.81	2.96	2.88	2.79	2.75
Anxiety	3.43	2.99	2.59	3.03	2.97	3.16	2.66
Dependence	3.13	3.05	2.93	3.00	2.89	3.11	2.88
Sentimentality	3.46	3.24	3.41	3.09	3.20	3.68	3.08
Social Self-Esteem	3.73	3.23	3.24	3.25	3.36	4.14	3.50
Social Boldness	3.07	3.05	2.83	2.98	2.91	3.28	2.99
Sociability	3.34	3.00	2.88	3.00	2.78	2.71	2.70
Liveliness	3.51	3.22	2.98	3.03	2.98	3.25	2.77
Forgiveness	2.83	3.04	3.06	2.94	2.80	2.90	2.98
Gentleness	3.26	3.40	3.49	3.03	3.14	3.52	3.26
Flexibility	3.10	3.12	2.84	2.98	2.71	2.81	2.91
Patience	3.26	3.04	3.30	3.02	3.23	3.71	3.52
Organization	3.57	3.35	3.06	3.04	3.19	3.11	2.88
Diligence	3.90	3.73	3.09	3.22	3.22	3.81	3.06

Perfectionism	3.56	3.41	3.16	3.13	3.31	3.48	2.95
Prudence	3.49	3.37	3.07	3.02	3.21	3.12	2.95
Aesthetic Appreciation	3.53	3.18	2.95	3.04	3.06	2.88	2.82
Inquisitiveness	3.51	3.20	3.01	3.01	3.11	2.72	2.74
Creativity	3.60	3.41	3.15	3.24	3.10	3.36	2.91
Unconventionality	3.54	3.38	3.05	3.10	3.00	3.86	3.10
Altruism	3.91	4.01	3.92	3.36	3.61	4.25	3.61

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Table 34

Facet Level Standard Deviation for HEXACO Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Sincerity	0.75	0.44	0.69	0.22	0.74	0.73	1.16
Fairness	0.83	0.44	0.96	0.74	1.11	0.65	1.40
Greed Avoidance	0.89	0.49	0.93	0.25	0.70	0.94	1.33
Modesty	0.73	0.32	0.62	0.25	0.70	0.89	1.20
Fearfulness	0.82	0.32	0.75	0.20	0.80	0.74	1.03
Anxiety	0.83	0.26	0.80	0.25	0.86	0.76	1.16
Dependence	0.81	0.32	0.78	0.16	0.77	0.71	1.01
Sentimentality	0.75	0.36	1.00	0.22	1.02	0.74	1.22
Social Self-Esteem	0.71	0.33	0.82	0.41	1.11	0.72	1.21
Social Boldness	0.78	0.28	0.89	0.29	0.96	0.89	1.19
Sociability	0.81	0.44	1.10	0.40	1.08	0.85	1.31
Liveliness	0.79	0.31	0.71	0.31	1.02	0.91	1.11
Forgiveness	0.81	0.26	0.80	0.15	0.92	0.68	1.16
Gentleness	0.68	0.38	0.86	0.16	0.99	0.69	1.08
Flexibility	0.70	0.35	0.82	0.14	0.89	0.75	1.14
Patience	0.81	0.20	0.52	0.16	0.93	0.70	1.09
Organization	0.83	0.36	0.93	0.22	0.84	0.85	1.10
Diligence	0.67	0.52	0.92	0.40	1.15	0.89	1.27

Perfectionism	0.71	0.43	0.75	0.34	0.97	0.87	1.13
Prudence	0.69	0.45	0.90	0.28	1.05	0.95	1.29
Aesthetic Appreciation	0.80	0.28	0.84	0.18	0.76	0.80	1.19
Inquisitiveness	0.82	0.31	0.82	0.10	0.69	0.66	1.03
Creativity	0.79	0.43	0.81	0.40	0.91	0.98	1.15
Unconventionality	0.61	0.35	0.80	0.21	0.76	0.60	1.12
Altruism	0.57	0.39	1.03	0.42	1.12	0.62	1.27

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Table 35

Domain Level Mean for HEXACO Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Hon	3.60	3.75	3.72	3.25	3.47	4.11	3.52
Emo	3.21	3.06	2.93	3.02	2.98	3.18	2.85
Ext	3.41	3.12	2.98	3.06	3.01	3.35	2.99
Agr	3.11	3.15	3.17	2.99	2.97	3.23	3.17
Con	3.63	3.47	3.09	3.10	3.23	3.38	2.96
Ope	3.54	3.29	3.04	3.10	3.07	3.20	2.89

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience.

Table 36

Domain Level Standard Deviation for HEXACO Human Responses and LLM-Generated Responses

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Hon	0.56	0.27	0.62	0.28	0.68	0.66	1.18
Emo	0.57	0.21	0.58	0.13	0.66	0.56	0.87
Ext	0.60	0.24	0.78	0.27	0.95	0.71	1.07
Agr	0.57	0.22	0.66	0.11	0.86	0.61	1.01
Con	0.53	0.33	0.68	0.25	0.83	0.75	1.00
Ope	0.55	0.23	0.75	0.15	0.72	0.58	1.01

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience.

Table 37

Standardized Factor Loadings for HEXACO Four-Factor Models of Each Domain

		human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
	Sincerity=~item6	.53	.41	.89	.25	.80	.62	.83
	Sincerity=~item30	.54	-.18	.34	.49	.14	.70	.78
	Sincerity=~item54	.67	.46	.70	.48	.92	.36	.65
	Sincerity=~item78	.56	-.17	.11	.86	.07	.81	.71
	Fairness=~item12	.73	.54	.81	.83	.92	.88	.94
	Fairness=~item36	.62	.68	.61	.66	.93	.85	.93
	Fairness=~item60	.50	.23	.58	.81	.49	.67	.78
	Fairness=~item84	.80	.41	.69	.64	.97	.89	.96
	Greed Avoidance=~item18	.54	.05	.25	.56	.26	.67	.74
	Greed Avoidance=~item42	.73	.51	.75	.66	-.94	.80	.86
	Greed Avoidance=~item66	.76	.61	.85	.77	-.94	.86	.93
	Greed Avoidance=~item90	.77	.78	.89	.67	-.96	.90	.93
	Modesty=~item24	.43	.18	.38	.50	.31	.73	.50
	Modesty=~item48	.47	.08	-.28	.34	-.09	.63	.64
	Modesty=~item72	.59	.55	-.80	.53	.87	.87	.94
Hon	Modesty=~item96	.82	.60	-.83	.76	.96	.87	.94
	Fearfulness=~item5	.58	-1.16	.94	.15	.82	.46	.92
	Fearfulness=~item29	.62	.04	.28	.81	.74	.50	.29

Emo

	Fearfulness=~item53	.70	-.12	.87	.65	.88	.70	.93
	Fearfulness=~item77	.54	.14	.09	.47	.55	.58	.46
	Anxiety=~item11	.72	.67	.85	.84	.88	.72	.89
	Anxiety=~item35	.70	.19	.32	.67	.72	.70	.76
	Anxiety=~item59	.51	.27	.19	.59	.57	.40	.65
	Anxiety=~item83	.61	.65	.86	.76	.89	.68	.88
	Dependence=~item17	.74	.55	.66	.42	.70	.82	.65
	Dependence=~item41	.63	.05	.51	.16	.03	.26	.59
	Dependence=~item65	.65	.47	.60	.84	.95	.70	.56
	Dependence=~item89	.50	.07	.27	.54	.69	.30	.56
	Sentimentality=~item23	.67	.54	.81	.36	.85	.62	.77
	Sentimentality=~item47	.54	.70	.90	.52	.89	.85	.82
	Sentimentality=~item71	.55	.56	.80	.76	.94	.83	.91
	Sentimentality=~item95	.61	.18	.71	.55	.72	.42	.84
	Social Self-Esteem=~item4	.66	.79	.54	.73	.85	.87	.87
	Social Self-Esteem=~item28	.44	.61	.90	.60	.88	.70	.75
	Social Self-Esteem=~item52	.65	-.21	.60	.58	.83	.79	.92
	Social Self-Esteem=~item76	.70	.17	.64	.53	.83	.80	.89
	Social Boldness=~item10	.58	.12	.70	.69	.63	.73	.70
	Social Boldness=~item34	.74	.67	.87	.92	.90	.81	.82
	Social Boldness=~item58	.66	.73	.51	.93	.87	.73	.69
	Social Boldness=~item82	.44	.08	.70	.67	.73	.61	.73
Ext	Sociability=~item16	.55	.41	.72	.86	.87	.67	.81

	Sociability=~item40	.73	.75	.84	.87	.98	.74	.90
	Sociability=~item64	.69	.63	.81	.78	.95	.72	.87
	Sociability=~item88	.67	.58	.89	.90	.93	.59	.86
	Liveliness=~item22	.65	.50	.80	.54	.88	.55	.65
	Liveliness=~item46	.76	.83	.88	.73	.86	.87	.83
	Liveliness=~item70	.66	-.58	-.29	.58	.76	.72	.62
	Liveliness=~item94	.70	-.26	.37	.73	.74	.55	.66
	Forgiveness=~item3	.72	.36	.69	.69	.82	.72	.78
	Forgiveness=~item27	.71	.33	.65	.35	.96	.65	.62
	Forgiveness=~item51	.51	.42	.75	-.37	.64	.40	.87
	Forgiveness=~item75	.74	.40	.66	-.11	.84	.46	.87
	Gentleness=~item9	.58	.34	.63	.37	.76	.48	.33
	Gentleness=~item33	.60	.48	.86	.71	.94	.72	.90
	Gentleness=~item57	.50	.62	.78	.73	.92	.70	.87
	Gentleness=~item81	.49	.75	.70	.68	.59	.73	.77
	Flexibility=~item15	.49	.26	.78	.51	.82	.48	.40
	Flexibility=~item39	.45	.56	.74	.81	.62	.62	.81
	Flexibility=~item63	.61	.17	.82	.86	.86	.76	.87
	Flexibility=~item87	.61	.35	.68	.90	.91	.59	.85
	Patience=~item21	.68	.06	.63	.14	.93	.80	.84
	Patience=~item45	.65	.30	.23	1.03	.69	.64	.73
	Patience=~item69	.64	.50	-.17	.00	.56	.34	.45
Agr	Patience=~item93	.70	.40	.90	.76	.89	.77	.89

	Organization=~item2	.57	.47	.58	.40	.88	.66	.77
	Organization=~item26	.64	.79	.71	.90	.72	.89	.88
	Organization=~item50	.64	.23	.54	.32	.42	.37	.59
	Organization=~item74	.72	.39	.60	.31	.71	.52	.55
	Diligence=~item8	.53	.78	.75	.87	.90	.87	.74
	Diligence=~item32	.68	.76	.89	.66	.87	.81	.59
	Diligence=~item56	.64	.34	.24	.45	.75	.57	.82
	Diligence=~item80	.57	.19	.12	.59	.82	.70	.86
	Perfectionism=~item14	.55	.58	.78	.85	.86	.79	.72
	Perfectionism=~item38	.61	.30	.39	.63	.68	.66	.83
	Perfectionism=~item62	.61	.72	.86	.69	.87	.85	.90
	Perfectionism=~item86	.57	.30	.39	.64	.79	.60	.40
	Prudence=~item20	.60	.46	.78	.59	.98	.88	.87
	Prudence=~item44	.69	.43	.62	.62	.76	.60	.55
	Prudence=~item68	.50	.57	.43	.65	.39	.81	.87
Con	Prudence=~item92	.55	.47	.83	.73	.79	.80	.83
	Aesthetic Appreciation=~item1	.66	.53	.71	.90	.70	.66	.93
	Aesthetic Appreciation=~item25	.54	.57	.67	.13	.76	.66	.86
	Aesthetic Appreciation=~item49	.64	-.40	.75	.10	.81	.53	.76
	Aesthetic Appreciation=~item73	.43	-.08	.27	.30	.64	.39	.44
	Inquisitiveness=~item7	.62	-2.52	.72	1.36	.94	.67	.95
	Inquisitiveness=~item31	.54	-.19	.58	.34	.80	.66	.76
	Inquisitiveness=~item55	.62	.05	.72	.00	.56	.53	.91

Ope

	Inquisitiveness=~item79	.64	-.01	.71	-.02	.50	.19	.29
	Creativity=~item13	.45	.40	.73	.75	.64	.46	.72
	Creativity=~item37	.73	.58	.68	.83	.93	.81	.80
	Creativity=~item61	.57	.69	.66	.37	.96	.76	.89
	Creativity=~item85	.69	-.04	.56	.89	.57	.84	.70
	Unconventionality=~item19	.34	.19	.72	.65	.59	.30	.84
	Unconventionality=~item43	.46	.73	.73	.60	.99	.89	.82
	Unconventionality=~item67	.37	.52	.19	.31	.68	.76	.41
	Unconventionality=~item91	.63	.29	.73	.56	.52	.25	.90

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.

Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience.

Table 38

Standardized Factor Loadings for HEXACO Six-Factor Model

	human responses	*persona* GPT3.5	*shape* GPT3.5	*persona* GPT4	*shape* GPT4	*persona* LLaMA3	*shape* LLaMA3
Hon=~Sincerity	.52	.57	.65	NA	.75	.76	.93
Hon=~Fairness	.58	.34	.52	NA	.69	.59	.83
Hon=~Greed Avoidance	.63	.36	.86	NA	.86	.78	.91
Hon=~Modesty	.52	.53	.77	NA	.85	.86	.95
Emo=~Fearfulness	.58	.32	.99	NA	.95	.33	.87
Emo=~Anxiety	.80	.35	.67	NA	.85	.31	.93
Emo=~Dependence	.38	.90	-.03	NA	.19	.69	.29
Emo=~Sentimentality	.46	.42	-.20	NA	.15	1.02	.47
Ext=~Social Self-Esteem	.77	.68	.89	NA	.89	.87	.95
Ext=~Social Boldness	.58	.59	.90	NA	.88	.72	.84
Ext=~Sociability	.54	.53	.80	NA	.79	.68	.68
Ext=~Liveliness	.83	.52	.78	NA	.96	.84	.86
Agr=~Forgiveness	.66	.70	.93	NA	.91	.81	.85
Agr=~Gentleness	.65	.58	.78	NA	.96	.80	.84
Agr=~Flexibility	.65	.59	.83	NA	.86	.82	.90
Agr=~Patience	.72	.53	.73	NA	.85	.83	.87
Con=~Organization	.65	.68	.90	NA	.89	.81	.89
Con=~Diligence	.71	.61	.60	NA	.88	.77	.64

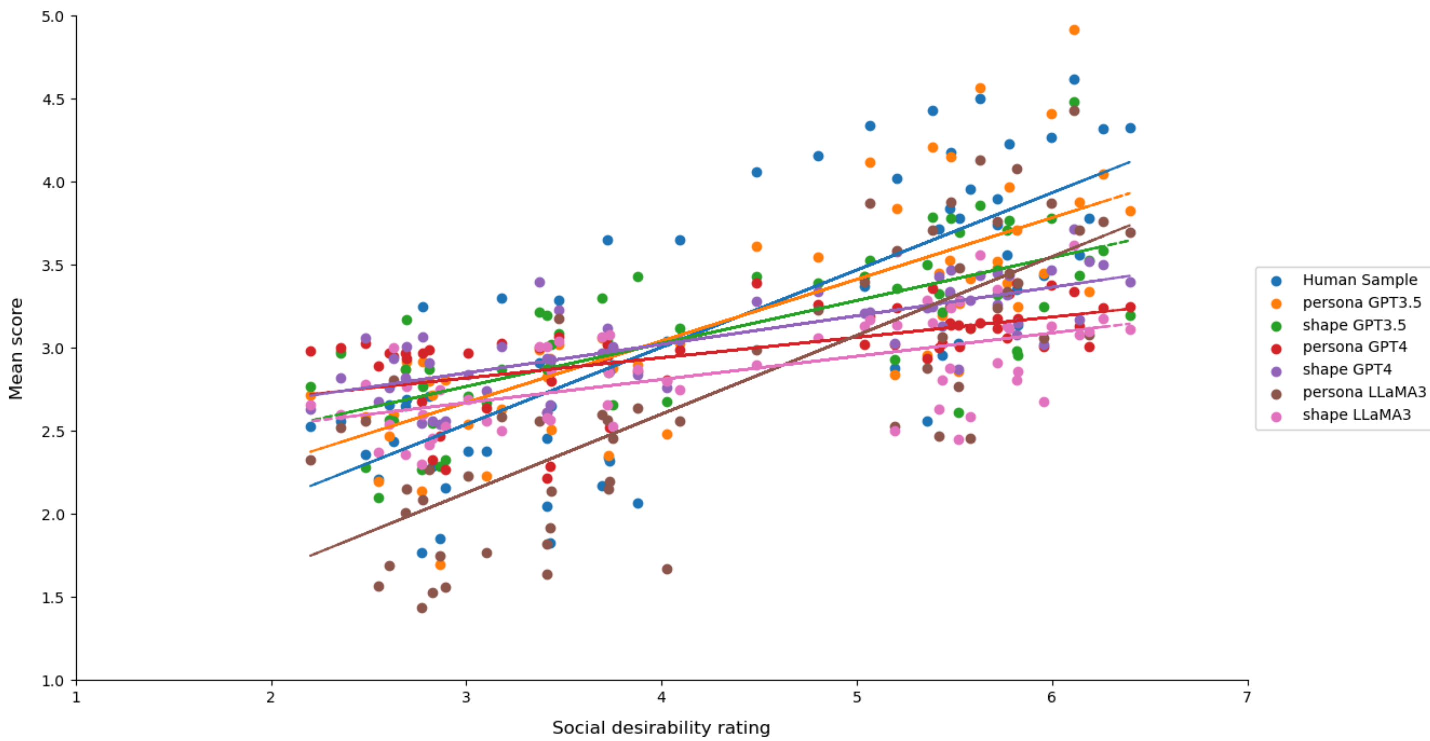
Con=~Perfectionism	.51	.59	.64	NA	.88	.80	.76
Con=~Prudence	.59	.71	.64	NA	.45	.79	.84
Ope=~Aesthetic Appreciation	.71	.35	.91	NA	.93	.90	.92
Ope=~Inquisitiveness	.50	.49	.91	NA	.88	.62	.92
Ope=~Creativity	.66	.67	.84	NA	.92	.55	.75
Ope=~Unconventionality	.59	.58	.86	NA	.87	.46	.85

Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses. The model did not converge for data generated by the GPT4 and *persona* method combination. Hon = Honesty-Humility; Emo = Emotionality; Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Ope = Openness to Experience.

Social Desirability Rating

Figure 1

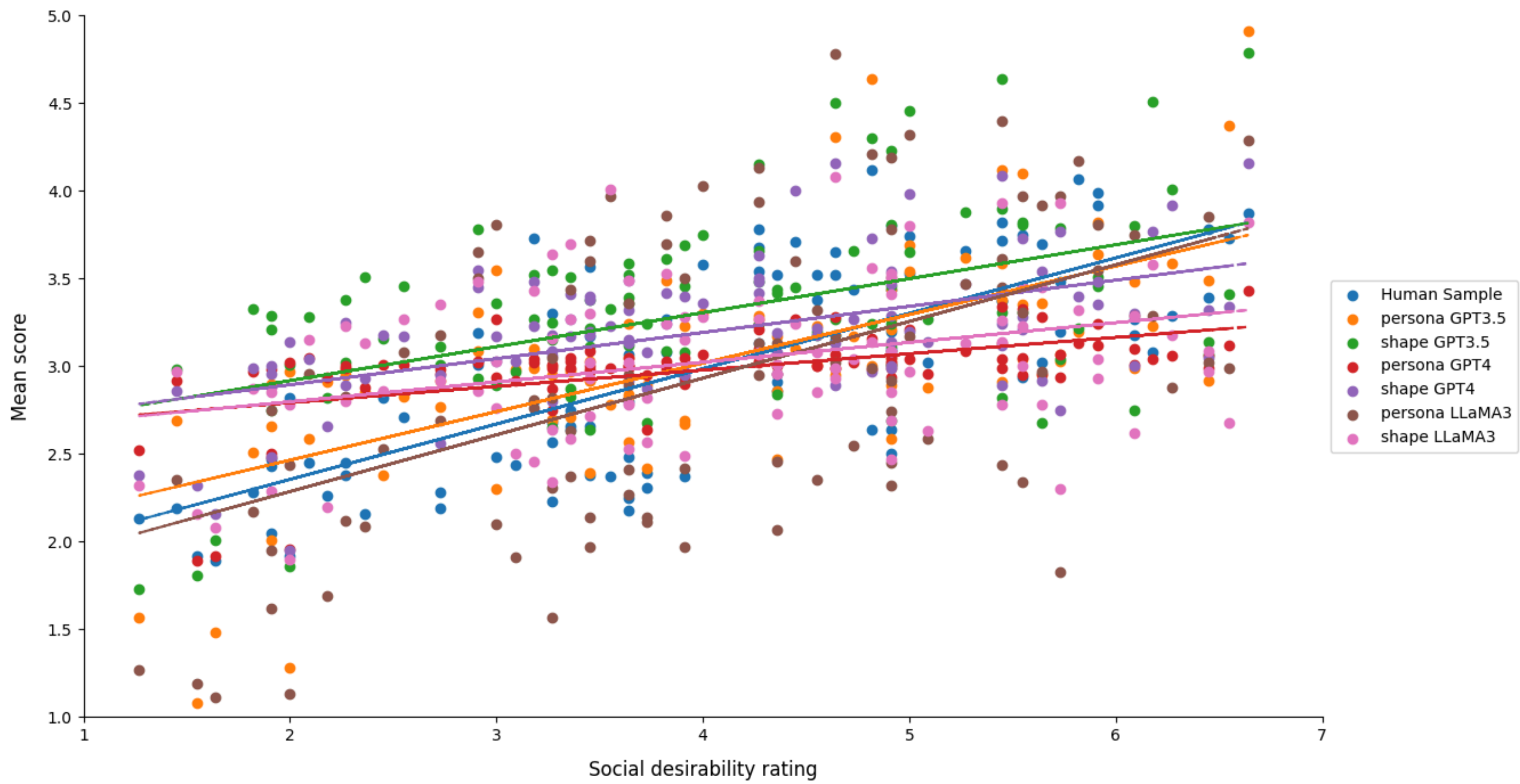
Regression Line Chart of Social Desirability Ratings: Human Responses vs. BFI-2 LLM-Generated Responses



Note. n = 1,559 for human responses, n = 299 for *persona* GPT-3.5, n = 297 for *shape* LLaMA-3, and n = 300 for other LLM-generated responses.

Figure 2

Regression Line Chart of Social Desirability Ratings: Human Responses vs. HEXACO-100 LLM-Generated Responses



Note. n = 7,204 for human responses, n = 298 for *persona* GPT-3.5, and n = 300 for other LLM-generated responses.